Department of Electrical and Computer Engineering

An Online Solution for Localisation, Tracking and Separation of Moving Speech Sources

Nicholas Chong Ewe Hai

This thesis is presented for the Degree of

Doctor of Philosophy

of

Curtin University

7th November 2015

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: . . . . . .

Date: . . . . . .

Acknowledgements

First and foremost, I would like to thank God for giving me the strength and discipline to finish this thesis. I believe it is through the hand of fate that I embark upon this journey of pursuing a PhD. For all the times when I feel my passion for research waver, it is my belief in God and His plans for me to do something good with my research that helps reignite this passion.

I would also like to express my gratitude to my supervisors, Prof. Sven Nordholm, Dr. Iain Murray and A/Prof. Ba Tuong. Throughout this journey, you all have provided me with more than just technical guidance. I sincerely thank you all for the moral support that you have given and the philosophies you have shared. It has been a long and arduous journey and I am grateful for your patience in guiding me. Furthermore, I would also like to thank my colleagues in the department for all the fruitful discussions and collaborations we shared.

Last but not least, I would like to express my deepest appreciation to my family and friends who have accompanied me throughout this journey. I am unable to list all the names but special mention goes to - my parents, Chong Thien Khee and Irene Liew, who always believe in me even when I doubt myself; my godmums, Liew Su Jin and Liew Sze Kian, who always take good care of me whenever I take a break from this journey; my brother, Kelvin Chong, who helps take care of the family while I am gone; my best friends, Seng Kiat, Pei Chee, Wee Lih and Ngie Chang - with friends like these, I do not need enemies; my girlfriend Poi, who threatened to break up with me if I do not finish this thesis by the end of this month.

Abstract

The sound source separation problem has challenged researchers in the field of acoustics for decades. Advances have been made in the field of blind source separation and various methods such as Independent Component Analysis, Higher Order Statistics and Blind Source Signal (BSS) separation via signal sparsity have been developed to separate non-moving sound sources. These BSS methods work by estimating the mixing parameters of the sound source mixture and demixing the sound sources by attributing the correct parameters to their respective sources. For non-moving speech sources, these mixing parameters and the number of sound sources remain constant throughout the whole period of measurement. However, non-moving sound sources do not fully represent the real world scenario.

A more complex scenario that better reflects the real world are sound signals which are dynamic in motion. Furthermore, the number of sound sources is not known *a priori* and the number of speech sources which are active remains unknown as these speech sources appear and disappear throughout the measurement period. Traditional BSS methods require the number of sound sources to be constant throughout the measurement period so a time varying number of speech sources is beyond the limitations of traditional BSS methods. Moreover, the challenge presented by this problem is not just the time varying number of speech sources. The mixing parameters of moving speech sources also change with time. As a result of this, traditional BSS methods are limited by their capability to properly attribute the correct mixing parameters to their respective sources.

The problem of localising, tracking and separating a time varying number of moving speech sources remains a challenging problem. Here, we are proposing an alternative approach by breaking this complex problem down into several problems which are solved in different stages - source localisation, source tracking and source separation.

Each of these stages considers various aspects of the source separation problem so a solution which takes all these aspects into consideration should be able to solve the source separation problem in a more complete way.

In this thesis, a robust online solution which localises, tracks and separates multiple moving speech sources in a room environment is proposed. The proposed solution uses multiple microphone arrays with an arbitrary geometry to record the sound sources. The signal sparsity property in the recorded source mixture is exploited to extract the acoustic features. The acoustic features from the multiple arrays are fused using a Random Finite Set (RFS) method in order to localise and track the moving speech sources. The time varying number of speech sources is also estimated using the RFS method. A track management extension is used to provide the labels required in the construction of the time frequency masks. The signal sparsity of the source mixture is once again exploited by applying the time frequency masks to separate the source mixture. The outlined method is capable of estimating the time-varying number of speech sources as well as tracking and separating them. It is also capable of mitigating the effects of noise and reverberation on sound source separation.

Contents

List of Figures

List of Algorithms

List of Terms and Acronyms

IID	Interaural Intensity Difference
IPD	Interaural Phase Difference
TF	Time Frequency
BSS	Blind Source Separation
HOS	Higher Order Statistics
SOS	Second Order Statistics
CASA	Computer Auditory Scene Analysis
NOSET	Number of Source Estimation Technique
RFS	Random Finite Set
DoA	Direction of Arrival
STFT	Short Time Fourier Transform
TDoA	Time Difference of Arrival
GCC	Generalised Cross-Correlation
MUSIC	MUltiple SIgnal Classification
SRP-PHAT	Steered Response Power with Phase Transform
ML	Maximum Likelihood
SCOT	Smoothed Coherence Transform
SMC	Sequential Monte Carlo
PF	Particle Filter
ML	Maximum Likelihood
MHT	Multiple Hypothesis Tracker
JPDA	Joint Probabilistic Data Association
PHD	Probability Hypothesis Density
CPHD	Cardinalized Probability Hypothesis Density
CBMeMBer	Cardinality Balanced Multi-target Multi-Bernoulli

MAST	Multiple Acoustic Source Tracking
DUET	Degenerate Unmixing Estimation Technique
DESPRIT	DUET ESPRIT
MENUET	Multiple Sensor DUET
OSPA	Optimatal SubPattern Assignment
ISR	Image-to-Source Distortion Ratio
SIR	Source-to-Interference Ratio
SAR	Source-to-Artifact Ratio
SDR	Signal-to-Distortion Ratio
MOS	Mean Opinion Score
VAD	Voice Activity Detector

List of Mathematical Notations

$\mathbf{Y}$	Matrix of received microphone signals
$\mathbf{A}$	Matrix of mixing parameters for the sound source mixture
$\mathbf{S}$	Matrix of sound source signals
M	Number of received microphone signals
K	Number of time frames
$y_M(t_K)$	Received signal from the M^{th} microphone in the K^{th} time frame
N	Number of sound sources
$s_N(t_K)$	N^{th} sound source in the K^{th} time frame
a_{mn}	Attenuation between the n^{th} sound source and the m^{th} microphone
h_{mn}	Impulse response between the n^{th} sound source and the m^{th} microphone
b	Room filter length
B	Maximum Room filter length
δ_{mn}	Delay between the n^{th} sound source and the m^{th} microphone
$a_{mn,x_{t,n}}$	Attenuation between the n^{th} sound source and the m^{th} microphone depending on the state of sound source at time t, $x_{t,n}$
$\delta_{mn,x_{t,n}}$	Delay between the n^{th} sound source and the m^{th} microphone depending on the state of sound source at time t, $x_{t,n}$
$x_{t,n}$	State of the n^{th} sound source at time t
$\mathfrak{p}_{t,n,\mathrm{x_{coord}}}$	x coordinate of the n^{th} sound source at time t
$\mathfrak{p}_{t,n,\mathrm{y_{coord}}}$	y coordinate of the n^{th} sound source at time t
$\dot{\mathfrak{p}}_{t,n,\mathrm{x_{coord}}}$	x velocity of the n^{th} sound source at time t
$\dot{\mathfrak{p}}_{t,n,\mathrm{y_{coord}}}$	y velocity of the n^{th} sound source at time t
$R_{yy}(\tau,\omega)$	The cross power spectrum of the STFTed signal, Y
$Y(\tau,\omega)$	The STFTed signal

τ	Time notation in the time frequency domain
ω	Frequency notation in the time frequency domain
$\mathbf{E}[\cdot]$	Expected value operator
W	Unmixing matrix
R_{xx}	Covariance matrix
J	The cost function
$\Lambda_y(\tau, \omega)$	The estimate of the cross-power spectrum of the model sources
v	Additive noise in a signal
$R_{\text{gcc}}(\delta)$	Generalised Cross Correlation function
$\mathcal{S}_{y_1,y_2}(\omega)$	Cross spectral density of the signals $y_1(t)$ and $y_2(t)$
X	A RFS set containing the states of the sound sources, x
$\hat{\theta}$	Estimated DoA
$\hat{\delta}_{\text{true},k}$	Estimated TDoA at time k
c	Speed of sound
d	Distance between microphone pair
$\mathbf{y}(\omega)$	Vector containing the narrowband received signals, $y(\omega)$
$\mathbf{A}_{\text{steering}}$	A matrix of the steering vector, $\mathbf{a}(\theta)$
$\mathbf{s}(\omega)$	Vector containing the narrowband source signals, $s(\omega)$
$\mathbf{v}(\omega)$	Vector containing the narrowband additive noise, $v(\omega)$
$P_{\text{MUSIC}}(\theta)$	The MUSIC spectrum
$a(\theta)$	The steering vector in MUSIC algorithm
$[\cdot]^H$	Hermitian transpose
μ_n	Decomposed eigenvalue in MUSIC algorithm
$\mathcal{P}(\mathfrak{p})$	Steered response power
$G_n(\omega)$	SRP-PHAT filter
$\mathfrak{p}$	Position of a sound source
$\hat{\mathfrak{p}}_{s,k}$	Estimated position of a sound source
x_k	Single target state at time k
$\mathcal{X}$	State space
$\mathcal{Z}$	Observation space
$f_{k\|k-1}(x_k\|x_{k-1})$	Transition density, also known as dynamics model
$g(z_k\|x_k)$	Likelihood function

$p_k(x_k\|z_{1:k})$	Posterior density
$\{\mathscr{X}_0^{(l)}\}_{l=1}^{L}$	A set of particles
$\{w_0^{(l)}\}_{l=1}^{L}$	A set of particle weights
$\mathcal{X}$	State space
$\mathcal{Z}$	Observation space
X_k	Multi-target state at time k
Z_k	Multi-target observation at time k
$\mathcal{F}(\mathcal{X})$	Finite subsets of the state space, $\mathcal{X} \subseteq \mathbb{R}^{n_x}$
$\mathcal{F}(\mathcal{Z})$	Finite subsets of the observation space, $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$
$p_{S,k}(x_{k-1})$	Survival probability
$f_{k\|k-1}(x_k\|x_{k-1})$	Transition density
$S_{k\|k-1}(x_k - 1)$	Survival RFS
Γ_k	Multi-target birth model
$p_{D,k}(x_k)$	Probability of detection
$\Theta_k(x)$	Measurement RFS
K_k	Clutter RFS
$\kappa_k(\cdot)$	Intensity function
$\pi_k(\cdot\|Z_{1:k})$	Multi-target Bayes posterior density
$f_{k\|k-1}(\cdot\|\cdot)$	Multi-target transition density
$g_k(\cdot\|\cdot)$	Multi-target likelihood
δX	Multi-target set integral
$\mathrm{MEL}(f_l)$	MEL frequency band
f_l	Linear frequency
$\tilde{R}_{gcc}(\delta)$	Clipped likelihood function in MAST
$H(\tau, u, \mathbf{k})$	Response from the direction $\mathbf{k}$ in the u^{th} frequency
$\mathcal{P}(\tau, \mathbf{k})$	Marginal DoA likelihood
$\mathcal{H}_0$	The hypothesis that "No source is present"
$\mathcal{H}_1$	The hypothesis that "Source is present"
$\lambda(\tau)$	Likelihood ratio of MAST
λ_T	Threshold of the likelihood ratio
$\tilde{\lambda}(\tau)$	Modified likelihood ratio
γ_A	Attack constant in the Attack and Decay filter

γ_D	Decay constant in the Attack and Decay filter
$L(\tau, u, \theta_\tau^i, \phi_\tau^i)$	Spatial filter of MAST
W	Window function
$\mathcal{H}_{21}$	Acoustic transfer function
a	Instantaneous attenuation
d	Instantaneous delay
α	Symmetric attenuation
$\tilde{\alpha_n}$	Power weighted attenuation estimate
$\tilde{\delta_n}$	Power weighted delay estimate
Ω_n	Set of (τ, ω) points associated with the nth cluster)
$\mathfrak{v}$	Intensity function of a Poisson RFS
$\bar{N}$	Mean cardinality of Poisson RFS
r	Probability of the Bernoulli RFS being a unit set
p	Probability density distribution of the Bernoulli RFS
$\emptyset$	Empty set
$\mathbb{M}$	Number of independent Bernoulli RFS
$X^{(i)}$	Independent Bernoulli RFS
$\pi(X)$	Multi-Bernoulli probability density
$\pi_{k\|k-1}$	Predicted multi-target density
$r_{P,k\|k-1}^{(i)}$	Probability of existence for a state
$p_{P,k\|k-1}^{(i)}(x)$	Probability density of a state
$f_{k\|k-1}(\cdot\|\zeta)$	Single target transition density
$p_{S,k}(\zeta)$	Probability of the single target's survival
$\{(r_{\Gamma,k}^{(i)}, p_{\Gamma,k}^{(i)})\}_{i=1}^{\mathbb{M}_{\Gamma,k}}$	*Bernoulli components* of the birth RFS at time k
$\mathbb{M}_{k\|k-1}$	Number of predicted tracks
$\bar{N}_{k\|k-1}$	Mean cardinality of predicted multi-target state
π_k	CBMeMBer approximation of the multi-target posterior density
$r_{L,k}^{(i)}$	Legacy probability of existence for a state
$p_{L,k}^{(i)}$	Legacy probability density of a state
$r_{U,k}^{*}(z)$	Measurement updated probability of existence for a state
$p_{U,k}^{*}(x; z)$	Measurement updated probability density of a state
$\psi_{k,z}(x)$	Elevation of a sound source

$\mathbb{M}_{k\|k-1}$	Number of posterior hypothesized tracks
$\bar{N}_k$	Mean cardinality of posterior multi-target density
$w_{k-1}^{(i,j)}$	Weight of a particle in the SMC CBMeMBer
$\delta_{x_{k-1}^{(i,j)}}(x)$	Particle in the SMC CBMeMBer
$L_{k-1}^{(i)}$	Number of particles
$l_{P,k\|k-1}^{(i)}$	Predicted label
$\mathcal{T}_{k\|k-1}$	Prior track table
$l_{\Gamma,k}^{(i)}$	Spontaneous birth labels
$\mathcal{T}_k$	Updated track table
$l_{L,k}^{(i)}$	Legacy labels
$l_{U,k}(z)$	Measurement updated labels
$\mathfrak{p}$	Position vector of a sound source
$\bar{v}$	Steady state velocity
$B_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}}$	Steady rate constant
T_u	Time interval between each measurement update
$u_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}}$	Gaussian variable with zero mean and unit variance
F	Transition Matrix
Q	Variance Matrix
$\mathcal{N}$	Gaussian distribution
N_{ϖ}	Number of sources detected
Δ	TDoA of a particle
$\hat{\Delta}$	Mean TDoA observed by a sensor pair
$\tilde{\delta_{n,k}}$	Relative delay
$\hat{s}_n(t)$	Estimated n^{th} speech source
$s_{\mathrm{tar},n}(t)$	Original n^{th} speech source modified by some allowable distortions
f_{s}	Sampling frequency
q_0	Prior probability that none of the TDoA is due to the true source
q_{ϖ}	Probability that the ϖth TDoA resulted from the true source
σ_{Δ}	Standard deviation from the observed TDoA
r	Distance between sound source and microphone
$\hat{\delta_n}$	Estimated delay
$\hat{\alpha_n}$	Estimated attenuation

$\tilde{M}_n(\tau, \omega)$	Estimated time frequency mask
$\hat{S}_n(\tau, \omega)$	Estimated nth labelled sound source
$\bar{d}^{(c)}$	OSPA distance
p	OSPA error order
c	OSPA cut off
$\bar{e}^{c}_{p,\mathrm{loc}}$	pth order per target localisation error
$\bar{e}^{c}_{p,\mathrm{card}}$	pth order per target cardinality error
$\hat{s}_n(t)$	Estimated n^{th} speech source
$s_{\mathrm{tar},n}(t)$	Modified n^{th} speech source with some allowable distortions
$e_{\mathrm{int}}(t)$	The interference error term
$e_{\mathrm{art}}(t)$	The artifact error term
$e_{\mathrm{spat}}(t)$	The spatial error term

Chapter 1

Introduction

If you have the motivation, I don't think anything in this world is impossible.
First, you have to start moving. If you move, something will start.
– Kaito Daiki

1.1 Motivation

Humans are social animals and we rely heavily on audio cues in most of these social interactions. An important yet often overlooked skill that humans possess is the ability to pick up the sound source of interest among multiple active sound sources and discern the direction from which the sound source originates. Humans are capable of separating the sound source of interest from a sound mixture based on the received audio cues [1]. Humans can localise the sound source of interest based on the interaural intensity difference (IID) and the interaural phase difference (IPD) [2]. On the other hand, machines find this simple task of separating sound sources from a sound mixture very challenging. The sound source separation problem is the problem of using machines to separate sound sources from a sound source mixture. The motivation of this research lies in the use of inter-disciplinary engineering techniques to solve an ongoing problem in audio separation research namely the separation of a time varying number of moving speech sources.

The sound source separation problem is a difficult problem to solve in the acoustic research area as knowledge of the sound source mixtures are not available *a priori*. In real world scenarios, human speakers tend to take turns speaking and occasionally

move around. The number of speakers will appear and disappear over time as the human speakers take turns to speak and the state of the speakers will change over time as they move around. Thus, the proposed solution has to jointly estimate the number of speakers as well as the state of these speakers. The additional challenge is to estimate identity information to these speakers in order to track the speakers with respect to time. The emergence of blind source separation (BSS) techniques have gained prominence in recent years due to their capability in separating sound source mixtures without *a priori* knowledge of the sound source's mixing parameters. BSS techniques vary from the highly statistical dependent Higher Order Statistics (HOS) [3] to those which exploit sparseness of the sound mixtures in the time frequency (TF) domain [4]. The focus of conventional BSS techniques is the separation of non-moving sound sources from a sound source mixture.

Conventional BSS techniques are limited in their capabilities to track and separate a time varying number of moving speech sources due to several constraints. The first constraint is the estimation of the number of sound sources. Conventional BSS techniques such as HOS [5] and Second Order Statistics (SOS) [6] are only capable of separating sound sources when the number of microphones are equal to or more than the number of sound sources to be separated so such techniques are not applicable in this scenario where the number of sound sources are not known *a priori*. BSS via TF masking [4] is capable of separating the sound sources when the number of sound sources exceed the number of microphones by exploiting the sparseness property of the sound sources. However, this technique assumes the number of sound sources remain constant throughout the measurement period so there will be errors in the number estimated when the number of sound sources changes with time. The other constraint faced by conventional BSS techniques is the assumption that the mixing parameters of the source mixture do not change with time. In conventional BSS techniques, the sound sources are assumed to be non-moving so the mixing parameters of the sound source mixture also remain constant throughout the measurement period. When the sound sources move, their mixing parameters will change as well. Conventional BSS techniques which analyse the whole signal mixture are incapable of associating the correct mixing parameters to their respective individual sources when these mixing parameters change with time.

The novelty of this research is the solving of the limitations in current source separation techniques which are unable to estimate a time-varying number of moving sound sources as well as matching the correct acoustic features to their respective original sources. By employing the logic that humans use in localising and separating sound sources, the additional information gained from knowing the sound sources location and trajectory can provide valuable insights for sound source separation. Sound source separation for a more realistic acoustic scenario can be achieved with the information gained by knowing the sound sources' locations, identities and tracks.

The result of this research can be applied in various areas such as communication, machine interaction, home automation and even assistive technology for the vision impaired. An application of this technology in communication is video conferencing. In a conference room scenario, the speech information is communicated to other users who are geographically separated. Accordingly, there can be only a short delay from when the speech information is collected and the result is presented. This short delay implies that only online separation solutions can be considered. Due to this practical constraint, an offline algorithm which requires the whole set of input data to improve performance in terms of tracking and separation is not considered as a solution for the "conference room problem". [7] states that an online algorithm only has access up to the current data whereas an offline algorithm has access to future data. This view is reaffirmed in [8] and [9]. The difference between an online recursive solution and an offline solution is discussed in [10]: online recursive solutions in target tracking are only conditional on the measurements obtained before or at the current time step whereas the offline solution, also known as Bayesian smoothing, relies on having information on the whole set of measurements in order to improve tracking performance. With the emergence of virtual reality and electronic gaming technologies, source tracking and separation can provide a different method of interaction by using sound as a controlling input. Sound localisation has been shown to be useful in a home environment [11]. The additional capabilities of separating and identifying the individual sound sources will add context to the actions of the speakers in an indoor environment. This can be further expanded as an assistive technology to help people suffering from vision impairment navigate their way indoors. Given the various challenges presented and the significant impacts of source localisation, tracking, and separation, this

research is a worthwhile venture.

1.2 Research Objectives

The main aim of this research is to localise, track and separate a time varying number of speech sources through the combination of sound source separation and RFS multi-target tracking techniques. A literature review is done to identify the main limitations of current source separation and multi-target tracking techniques in the applications of moving speech source tracking. The main aim can be achieved in three stages - acoustic feature extraction, target tracking and source separation. In terms of feature extraction, the research objectives are:

- To apply a suitable algorithm to extract audio features from an audio mixture
- To extract the audio features in underdetermined, determined and overdetermined scenarios

After the audio features have been extracted, we seek to achieve the following during the target tracking stage:

- To estimate the time varying number of speech sources based on the extracted audio features
- To estimate the state vectors of multiple moving speech sources and track these sources instead of just estimating their Direction of Arrival (DoA)
- To mitigate the effects of noise and reverberation on the separation of the sound source mixture

With the data available from target tracking, the research objectives to be met in the sound source separation stage are:

- To separate multiple moving speech sources based on tracking data
- To integrate existing methods of sound source separation with advance target tracking methods

- To investigate the differences in reconstructed sound quality between hard masking technique and soft masking technique and propose the better option for sound source reconstruction

In this thesis, we propose a new framework to solve the sound source separation problem for a time varying number of speech sources by overcoming the limitations of the conventional sound source separation techniques through an integration of the BSS technique with a principled Random Finite Set (RFS) [12] [13] technique. The problem of acoustic feature extraction and the subsequent separation can be overcome with the BSS technique while the problem of localisation, estimation of the source number as well as the problem of data association can be solved using the RFS algorithm.

1.3 Scope of the Research

The main focus of this thesis is to provide an recursive online solution to the source separation problem for a time varying number of moving speech sources. Hence, the scope of this research covers sound source separation, sound source localisation and tracking. Real life audio separation is very demanding so the audio scenarios investigated will be more complicated: First of all, the number of speech sources available in the sound mixture will be unknown to the system and the number of speech sources appearing in the scenarios will vary with time. Secondly, the speech sources are non-stationary and they will be moving with time. Thirdly, the identity association between the sound sources will be unknown to the system. Finally, the system will need to perform sound source separation based on the estimated data. To that end, the system will need to be able to extract audio features without *a priori* knowledge of the sound source mixture, estimate the number and state of these sound sources, track these sound sources' movements and separate the sound source mixture.

The secondary focus of this research focuses on using distributed microphone array for target tracking purposes. As the real world scenario is taken into consideration, the microphone arrays need to be able to extract the audio features when the number of sound sources are less than the microphones (overdetermined), equal to the number of microphones (determined) and more than the number of microphones (underdetermined). This research will look into fusing the readings from multiple microphone

pairs in order to enhance the accuracy of the target localisation and tracking. The data fusion will be done using a recursive Bayesian method. The geometry of the microphone array can be arbitrary as the RFS filtering technique performs the estimation on the targets' locations based on the best observable data.

The proposed technique is evaluated using simulated scenarios which closely mimic real world environments. Both noise and room reverberation are considered as sources of interference in our evaluations. As such, effects of noise and reverberation will be investigated and methods to mitigate the effects of noise and reverberation will be proposed.

1.4 Thesis Structure

The thesis is structured in such a manner that it mirrors the progress in the development of the proposed algorithm. The thesis is organized according to the following structure:

Chapter 2 outlines the scenario of the problem and the system model. This chapter also provides the background information on different methods available for sound source localisation, tracking and separation. Apart from that, a general overview on the theory behind these algorithms is given and the applicability of these algorithms in this research is discussed.

In Chapter 3, the motivation of using BSS vis signal sparsity approach to extract the audio features will be discussed. Apart from that, the theory behind the BSS approach used will also be explained in detail. The limitations of this original BSS technique will be examined and the reasons this technique is not suitable for tracking moving speech sources will be stated in this chapter as well.

Chapter 4 provides a detailed insight into the Random Finite Set (RFS) filtering method used to provide the target tracking capabilities for the proposed algorithm. A generic implementation of the proposed RFS filtering method will be discussed in Chapter 4 as well. This chapter also points out the drawback in the original RFS method which makes it not directly applicable to perform speech source separation. The proposed solution to this limitation will be detailed in Chapter 5.

Chapter 5 highlights the contribution of this thesis. This chapter discusses the adaptations made to the source separation technique and the RFS filtering technique in

order to apply them to the problem of separating a time varying number of moving speech sources. Both the single state dynamics model as well as the single state observation model used in the RFS filtering technique are also explained in this chapter. This chapter also contains information on the track management method used to provide labels to the tracked sound sources and the ways in which this identity information is subsequently used to construct the mask to perform sound source separation.

Chapter 6 discusses the numerical evaluations carried out to test the viability of the proposed solution. The effects of various environmental acoustic interference, such as noise and reverberation, on the accuracy of the sound source tracking and separation are investigated and discussed in this chapter.

Chapter 7 is the conclusion of the thesis with a summary of the contributions made. A brief discussion on the possible future work undertaken to further improve the proposed technique will also be provided in this chapter.

1.5 Thesis Contribution

The following constitutes the original contribution of this thesis:

- Identified the limitations that exist within the current approaches to the source separation problem for an unknown number of moving speech sources and proposed a solution that solves this problem as both an acoustic separation problem and a multi-target tracking problem. As the number of the speech sources are not known *a priori* and the speech sources are moving, the problem can be cast as a multi-target tracking problem in order to overcome the limitations of conventional source separation techniques in tracking moving speech sources.

- Proposed a recursive online solution to a more complex source separation problem that involves moving targets. Most of the conventional source separation techniques limit their scope to non-moving targets so the challenge is in separating moving sound sources. The complexity lies in identifying the correct relative delay and attenuation of each sound sources and associating these acoustic features with their respective sound sources as these acoustic features change with the movement of the sound sources. The proposed solution exploits the sparseness of the audio mixture in order to extract the acoustic features required by the

target tracking algorithm.

- Proposed a solution that integrates the existing sound source separation techniques with a principled target tracking approach in order to solve the problem of separating moving sound sources. The proposed solution fuses a BSS method with a RFS filtering method by using the RFS method to combine the spatio temporal information of the audio features with the dynamic model to provide a recursive Bayes optimal solution which is able to solve the sound source tracking and separation problem online.

- Most conventional sound source separation systems assume to have prior knowledge of the number of sound sources. The proposed method attempts to estimate the number of targets to be tracked based on the audio features extracted without *a priori* knowledge on the number of sound sources.

- Developed a soft masking technique for sound source reconstruction based on the estimated tracking results. A soft mask provides a better sound quality for the reconstructed signals as opposed to a hard mask.

1.6 Thesis Publication

The following publications are a direct result of the research work undertaken:

- **N. Chong**, S. Wong, B.T. Vo, S. Nordholm, and I. Murray, "Multiple sound source tracking via Degenerate Unmixing Estimation Technique and Cardinality Balanced Multi-target Multi-Bernoulli filter (DUET- CBMeMBer)," in *Proceedings of IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing, 2014. ISSNIP'14* IEEE, 2014, pp. 16.

- **N. Chong**, S. Wong, S. Nordholm, and I. Murray, "Multiple Sound Source Tracking and Identification via DUET CBMeMBer with Track Management," in *Proceedings of Asia Pacific Signal and Information Processing Association, 2014. APSIPA'14. Annual Summit and Conference*, Dec 2014, pp. 1-5.

- **N. Chong**, S. Nordholm, B.T. Vo, and I. Murray, "Tracking and Separation of Multiple Moving Speech Sources Via Cardinality Balanced Multi-Target Multi

Bernoulli (CBMeMBer) Filter and Time Frequency Masking" in *Proceedings of IEEE International Conference on Control, Automation and Information Sciences, 2016. ICCAIS'16* IEEE, Oct 2016, to be published.

A revised version of work featured in the publications is used in Chapter 3, Chapter 4, Chapter 5, and Chapter 6 of this thesis.

Chapter 2

Background

Even if your goal is noble, it is all meaningless if your methods are wrong!

– Kouta Kazuraba

This chapter will give the details of the problem that this research aims to solve in Section 2.1. The system model used in the research is introduced in Section 2.2. This is followed by a literature review on the different techniques in the field of sound sources separation, sound source localisation and target tracking as the final solution encompasses the techniques from these three research fields. The main theories as well as the suitability of these techniques will be discussed in Chapter 2. The literature review mirrors the development of this research. Hence, sound source separation techniques that extract audio features are reviewed first followed by sound source localisation, and target tracking techniques. An overview of the proposed solution is given in Section 2.6 after the literature review.

2.1 Problem Formulation

The problem this research aims to solve is a more realistic scenario of the source separation problem. In an acoustic scenario which is closer to the real world environment, the number of targets in a room is not known *a priori* and this number tends to change with time. Furthermore, the information that is available to the system is only up to the current time. The system has no access to future data as it has to process the acoustic signal as it is received. Moreover, these targets will not necessarily be stationary. The

targets may move in random patterns so a technique that aims to solve the more complex source separation problem has to account for these factors instead of just purely demixing the sound source mixture for stationary targets. An example of this problem in real life is the conference room scenario which is provided by figure 2.1.

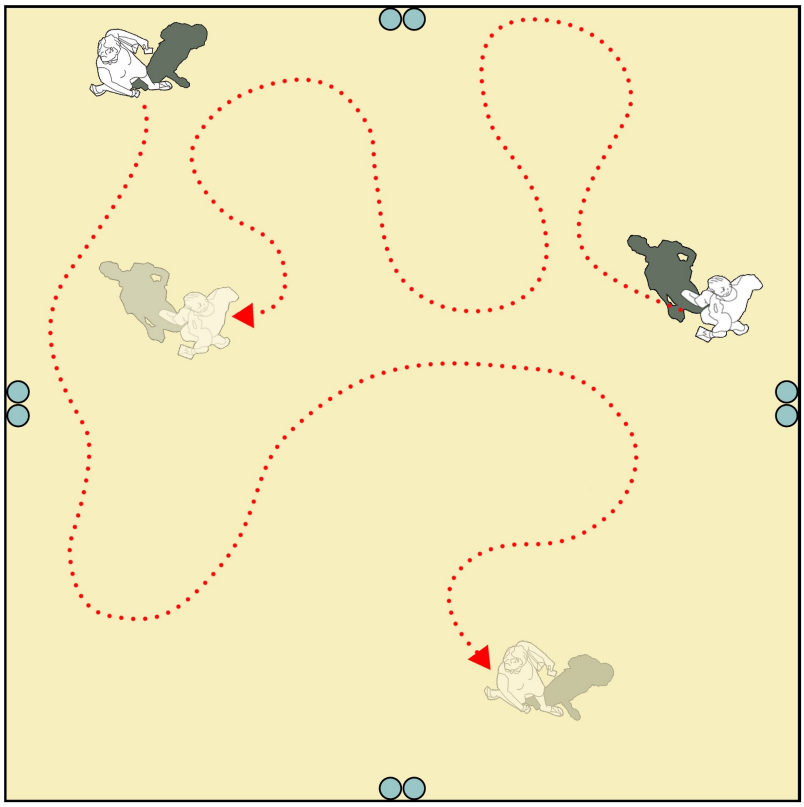


Figure 2.1: Typical source separation problem in a conference room scenario

In a "conference room scenario" whereby the number of acoustic targets vary with time and the targets are moving, the solution will need to be able to estimate the time varying number of targets, estimate the position of these targets as well as associate the identities of these targets in order to track them and separate them. Conventional BSS techniques are not applicable in the separation of a time varying number of sound sources as these techniques require the number of sound sources and the acoustic features of these sound sources to be constant throughout the whole measurement period in order to separate them. The "conference room problem" outlined is more than just a source separation problem as it requires the number and the state of the targets to be estimated before sound source separation can be performed. An additional challenge to the "conference room problem" is the need for the different speech sources to be tracked and separated as the acoustic signals are being received. Due to this practical

constraint, an online algorithm which is capable of processing the output as the sequence of acoustic data is being received is required. A system model that represents the acoustic signal model in a conference room is required before a solution can be proposed. Section 2.2 gives an overview of the general BSS and acoustic models prior to introducing a system model that fits the "conference room problem".

2.2 System Model

The sound source separation problem in a conference room scenario is a more complex form of the BSS problem. As a result of this, a basic understanding of the BSS problem is required in order to formulate the problem in the context of a "conference room scenario" with a time varying number of moving speech sources. The general BSS problem can be summarised as a problem of finding the mixing parameters of a sound source mixture and inverting the received signals using the estimated mixing parameters in order to separate the sound sources. The BSS problem can be mathematically expressed as

$$\mathbf{Y} = \mathbf{A} \circ \mathbf{S} \tag{2.1}$$

whereby $\circ$ denotes a multiplication or a convolution depending on the acoustic signal model used. $\mathbf{Y}$ refers to an $M \times K$ matrix of received signals

$$\mathbf{Y} = \begin{bmatrix} y_1(t_1) & y_1(t_2) & \cdots & y_1(t_K) \\ y_2(t_1) & y_2(t_2) & \cdots & y_2(t_K) \\ \vdots & \vdots & \ddots & \vdots \\ y_M(t_1) & y_M(t_2) & \cdots & y_M(t_K) \end{bmatrix} \tag{2.2}$$

with M referring to the number of sound source signals received and K referring to the time frames. $\mathbf{S}$ is the $N \times K$ matrix of sound sources

$$\mathbf{S} = \begin{bmatrix} s_1(t_1) & s_1(t_2) & \cdots & s_1(t_K) \\ s_2(t_1) & s_2(t_2) & \cdots & s_2(t_K) \\ \vdots & \vdots & \ddots & \vdots \\ s_N(t_1) & s_N(t_2) & \cdots & s_N(t_K) \end{bmatrix} \tag{2.3}$$

with N being the number of sound sources. $\mathbf{A}$ is the $M \times N$ matrix of mixing parameters containing the acoustic features to be estimated using the BSS technique. The

acoustic features to be estimated and the difficulty in estimating these parameters rely on the acoustic signal model used for the BSS. In traditional BSS techniques, the size of **A** and **S** are rigid and constant throughout the measurement period.

There are several acoustic signal models which are commonly used in sound source separation - instantaneous, anechoic, and echoic or convolutive. The instantaneous acoustic signal model assumes that there is only an attenuation factor between the signals received by the microphones. The observed signal at the m^{th} microphone for the instantaneous acoustic model can be expressed as

$$y_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t) \tag{2.4}$$

with $y_m(t)$ being the received signal and $s_n(t)$ being the sound source while a_{mn} is the attenuation between the n^{th} sound source and the m^{th} microphone. The acoustic features in **A** which can be extracted from the instantaneous model is just a matrix of attenuation factors - $\{a_{11}, \cdots, a_{1N}, \cdots, a_{M1}, \cdots, a_{MN}\}$. This is the simplest acoustic model in the field of BSS research.

An anechoic audio mixing model is a more complex signal model that evolved from the instantaneous model. The anechoic acoustic model considers the attenuation and delay between the microphones. The anechoic signal model can be expressed as

$$y_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \delta_{mn}) \tag{2.5}$$

where δ_{mn} is the delay between the n^{th} sound source and the m^{th} microphone. The acoustic features in which are estimated by using the anechoic model are sets of attenuation factors and delays - $\{a_{11}, \cdots, a_{1N}, \cdots, a_{M1}, \cdots, a_{MN}\}$ and
$\{\delta_{11}, \cdots, \delta_{1N}, \cdots, \delta_{M1}, \cdots, \delta_{MN}\}$.

A convolutive acoustic signal model better represents real world signals than the instantaneous model and the anechoice model but it is also a more difficult acoustic model to solve as a BSS problem [14], [15]. It is expressed as a convolutive sum of the source's signal and the microphone's impulse response [16]

$$y_m(t) = \sum_{n=1}^{N} \sum_{b=0}^{B-1} h_{mn}(b) s_n(t - b) \tag{2.6}$$

with h_{mn} being the impulse response between the n^{th} sound source and the m^{th} microphone while b represents the room filter length and B is the maximum room filter

length. The acoustic features to be estimated using the convolutive acoustic model is a set of room impulse responses between all the N number of sources and the M number of microphones - $\{h_{11}, \cdots, h_{1N}, \cdots, h_{M1}, \cdots, h_{MN}\}$. When the sound sources are not moving, the impulse responses of the respective sound sources remain constant. Hence, the impulse response of the non-moving sound sources to be estimated throughout the measurement period will be the same.

In a conference room scenario, the number and the state of the sound sources are constantly changing. The conventional BSS model as well as the acoustic signal model used need to be able to accommodate these changing parameters. The number of sound sources, N, and the acoustic features in the mixing parameter matrix is now dependent on time. The size of the mixing parameter matrix will be $M \times N(t)$ and the size of the sound sources matrix will be $N(t) \times K$. As a result of this, the proposed solution has to be able to address the changes in the number of sound sources as well as the changes in the acoustic features. The convolutive signal model that represents this scenario is

$$y_m(t) = \sum_{n=1}^{N(t)} \sum_{b=0}^{B_n-1} h_{mn,x_{t,n}}(b) s_n(t - \delta_{mn,x_{t,n}}(b)). \tag{2.7}$$

The number of the active sound sources, $N(t)$, and the sound sources, $s_n(t-\delta_{mn,x_{t,n}}(b)$ are now changing with time. Consequently, the impulse response of these sound sources, will also be dependent on the state of these active sources. Hence, the problem of impulse response estimation which is already a difficult one will be even more challenging to solve.

In order to approximate the convolutive model of the "conference room problem", we propose to use the anechoic acoustic model which is a compromise between the instantaneous model and the convolutive model. The proposed anechoic signal model includes the parameters that account for the number of moving speech sources and the acoustic features which changes with time. The acoustic model proposed is

$$y_m(t) = \sum_{n=1}^{N(t)} a_{mn,x_{t,n}} s_n(t - \delta_{mn,x_{t,n}}) \tag{2.8}$$

with $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$ being the attenuation and delay of the n^{th} sound source which are now dependent on the state of the source, $x_{t,n}$, that changes with time. The state of a sound source can be defined to include the position and the velocity of the particular

sound source

$$x_{t,n} = \{\mathfrak{p}_{t,n,\mathrm{x_{coord}}}, \mathfrak{p}_{t,n,\mathrm{y_{coord}}}, \dot{\mathfrak{p}}_{t,n,\mathrm{x_{coord}}}, \dot{\mathfrak{p}}_{t,n,\mathrm{y_{coord}}}\}^T \tag{2.9}$$

with $\mathfrak{p}_{t,n,\mathrm{x_{coord}}}$ and $\mathfrak{p}_{t,n,\mathrm{y_{coord}}}$ representing the x and y coordinates respectively while $\dot{\mathfrak{p}}_{t,n,\mathrm{x_{coord}}}$ and $\dot{\mathfrak{p}}_{t,n,\mathrm{y_{coord}}}$ represents the velocities of the x and y. With the signal model established, we propose to break the "conference room problem" down into the following problems and systematically solving for each:

- Extraction of acoustic features, $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$ from a speech source mixture.
- Joint estimation of the speech sources' state, $x_{t,n}$, and number, $N(t)$.
- Calculation of the acoustic features, $\hat{a}_{mn,x_{t,n}}$ and $\hat{\delta}_{mn,x_{t,n}}$ based on the estimated state of the sound sources, $\hat{x}_{t,n}$.
- The problem of associating the estimated acoustic features, $\hat{a}_{mn,x_{t,n}}$ and $\hat{\delta}_{mn,x_{t,n}}$, with their respective speech sources based on the identity of the sources.
- The problem of speech source reconstruction based on the estimated acoustic features, $\hat{a}_{mn,x_{t,n}}$ and $\hat{\delta}_{mn,x_{t,n}}$.

As the individual speech sources are characterised by their set of acoustic features, $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$, it is physically intuitive to extract these features and track it recursively in order to localise and track the sound sources. These features also contain spatial information between the speech sources and the microphones. As the acoustic features are extracted when the acoustic signals are received by the microphones, only the current acoustic features are available to the system. This adds to the complexity of problem as the data used by the system for estimation is not available in its entirety. The state, $x_{t,n}$, and number, $N(t)$, of the speech sources can be tracked based on the observed acoustic features. The tracking of speech sources based on the acoustic features extracted is more mathematically tractable compared to using the raw information from the whole signal. If the raw information of the acoustic signal was used directly to estimate the number and the state of the speech sources, more sophisticated models would be required to represent the speech sources. This will lead to an increase in the computational cost of the solution. Once the state and the number

of the speech sources have been estimated, the estimated acoustic features, $\hat{a}_{mn,x_{t,n}}$ and $\hat{\delta}_{mn,x_{t,n}}$, which follow the speech sources trajectory can be obtained by exploiting the kinematic information on the speech sources movement. Through data association, estimated acoustic features which belong to the same speech sources can be grouped together. The speech sources can be separated and reconstructed based on the estimated acoustic features which are associated with them.

2.3 Sound Source Separation

The first that needs to be solved in this research is the extraction individual acoustic features, $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$ from a speech source mixture. BSS is a technique used to separate these sound mixtures without *a priori* information[17]. BSS separates the sound sources by estimating the mixing parameter which are usually the acoustic features of the individual sound sources. In essence, no prior knowledge about the sound mixture is required in order to achieve source separation though some BSS systems require certain assumptions to be made about the sound sources. A study on the BSS techniques will also provide clues on possible solutions to cluster the acoustic features of the sound sources and the reconstruction of these sound sources. In BSS, there are multiple methods to achieve source separation. The main branches of BSS are Higher Order Statistics (HOS), Second Order Statistics (SOS), Computational Auditory Scene Analysis (CASA) and Source Separation via Signal sparseness [15].

2.3.1 Higher Order Statistics

In HOS, the sound source model used is the instantaneous acoustic model in equation (2.4). HOS is one of the earliest BSS technique used so the system model used is $Y = AS$ as shown in equation (2.1) [18]. Based on the system model used, HOS assumes that the microphones receive M number of linearly independent sound sources and the transformation from the N number of sources to the received signals is a linear transformation that is invertible. Hence, the number of received signals M has to be greater than or equal to the number of sound sources N. In other words, HOS only works in determined or overdetermined conditions.

The underlying assumption in HOS is the statistical independence of the sound

sources. Sound sources are considered statistically independent if their joint probabilistic function is factorisable [19]. Another requirement imposed by Higher Order Statistics is the underlying sources need to be non-Gaussian [5]. In order to achieve source separation, the mutual dependence between the signal models are minimized [15]. Independent Component Analysis (ICA) [19] falls under the class of BSS using HOS. In ICA, the aim is to maximize the independence of the signals and this can be achieved through the minimisation the mutual information between two sources [5], maximization of entrophy [20] or non-Gaussianity [19]. The Information Theoretic criteria is used in ICA to minimize the mutual information between the sources by maximising the entrophy [15].

The advancement in ICA techniques have seen ICA being extended to other more complex signal models such as the convolutive model [21] [22]. As the convolutive model is more complex to solve than the instantaneous signal model, the ICA techniques that solves the convolutive model will introduce additional computational complexity. [22] simplifies the convolutive problem by solving it as an instantaneous problem at each of the frequency bins but it also introduces a frequency permutation ambiguity with this method [22]. Permutation alignment of each bin is performed to solve the permutation ambiguity problem [22]. Although ICA is mainly used for sound source separation, Sawada also used ICA as a sound source localisation algorithm in [23].

Despite the capability of ICA to extract acoustic features from a speech source mixture, ICA is not a suitable technique to extract acoustic features from a time varying number of moving speech sources due to several limitations. BSS via ICA suffers from both permutation ambiguity and scaling ambiguity problems [22] [24, p.29]. Due to the permutation ambiguity, the separated sound sources are not in the proper order [22]. The permutation ambiguity will be a more prominent issue when the acoustic features are to be associated with their respective moving sound sources. The separated sound sources cannot be properly scaled due to the scaling ambiguity. The sound sources are not properly scaled. For a time-varying number of speech sources, there is no *a priori* knowledge on the number of the speech sources. BSS via HOS which requires the number of received signals to be equal to or more than the number of sound sources is not applicable when the number of sound sources is not known beforehand

and that number is changing with time. Due to the nature in which ICA performs source separation by determining the independence of the underlying sound sources, the computational complexity involved is very high. The research aims to use a low computational complexity method that is capable of separating sound sources in over-determined, determined and underdetermined scenarios so ICA is deemed unsuitable as a feature extraction technique for this research.

2.3.2 Second Order Statistics

The next BSS research that is explored is BSS using Second Order Statistics (SOS) [6]. SOS is the BSS technique that came after HOS. SOS only works in determined or overdetermined situations so the number of received signals, M, has to be more or equal to the number of sound sources, N. The signal property used by SOS to achieve source separation is the non-stationarity of the signals. SOS has various advantages over HOS. One of such advantages is the computational complexity. With the assumption of non-stationarity, it is capable of performing BSS with less data and computational complexity [25]. SOS is also more capable in handling Gaussian signals as long as the signals are non-stationary or coloured [25].

This is notably useful as real world signals are non-stationary most of the time and this condition imposed upon the signals can be easily fulfilled [25]. Speech signal fulfills the condition of non-stationarity so SOS can be applied to separate speech mixtures. The covariances of non-stationary sources are linearly independent at different time intervals. The temporal varying statistics of these non-stationary sources provide additional information needed to achieve source separation. SOS generally performs decorrelation which does not necessarily indicate that the sound sources are independent though independent sources are definitely decorrelated [25]. With the additional information available at varying time intervals, decorrelation can be performed by minimizing cross power estimates [26].

Simultaneous minimization of multiple cross correlations is required in order to have enough second order constraints to separate the signals [15]. A frequency domain implementation of simultaneous decorrelation diagonalization of the cross power spectrum was proposed by Parra and Spence [27]. The cross power spectrum can be

defined as

$$\begin{aligned} R_{yy}(\tau,\omega) &= \mathbf{E}[Y(\tau,\omega)Y^H(\tau,\omega)] \\ &= W(\omega)R_{xx}(\tau,\omega)W^H(\omega) \end{aligned} \tag{2.10}$$

with $R_{yy}(\tau,\omega)$ being the cross power spectrum and $\mathbf{E}[\cdot]$ being the expectation of the Short Time Fourier Transformed (STFT) signal, $Y(\tau,\omega)$, and its Hermitian transposed counterpart, $Y^H(\tau,\omega)$. W refers to the unmixing matrix and R_{xx} is the covariance matrix. The aim is to minimize the cross powers on the diagonals of the previous matrix by minimizing [15]

$$J = \sum_{\tau,\omega} ||R_{yy}(\tau,\omega) - \Lambda_y(\tau,\omega)||^2 \tag{2.11}$$

where J is the cost function and $\Lambda_y(\tau,\omega)$ is an estimate of the cross-power spectrum of the model sources which is assumed to be diagonal. This cost function captures various time and frequencies and is minimized with respect to $W(\omega)$ and $\Lambda_y(\tau,\omega)$. The independence conditions are fulfilled and the unmixing matrix, $W(\omega)$ can be determined if the signals are non-stationary and the cross powers differ for different time instances, τ.

Similar to HOS, SOS is mainly performed in the frequency domain so the convolutive acoustic model can be approximated by an instantaneous acoustic model. Hence, SOS also experiences the scaling and permutation ambiguities. The problems that arise from the permutation ambiguity will be even more evident when the acoustic features are to be matched with their original sources as the acoustic features of moving sources will change with time. Another reason SOS is not chosen to extract the acoustic features this research is the incapability of SOS to separate sound sources in underdetermined scenarios. As SOS relies on the assumption that the number of received signals, M, is equal to or greater than the number of sound sources, N, the maximum number of sound sources present has to be known *a priori* in order to allocate enough microphones. Due to these limitations, a different approach is required to extract the acoustic features of the moving speech sources.

2.3.3 Computer Auditory Scene Analysis

Both HOS and SOS are highly statistical BSS methods which estimate the acoustic features of sound sources in order to separate them. Moreover, both techniques rely

on strong assumptions which may not necessarily always hold true in real acoustic scenarios. Computer Auditory Scene Analysis (CASA) [28] takes a different approach to sound separation. CASA is a technique which models the human auditory system in artificial systems [29]. Unlike HOS or SOS, CASA does not rely on the strong signal assumptions so it is more robust in processing signals in real environments [29]. CASA has been shown to complement BSS techniques in the separation of sources in underdetermined scenarios [30] [31]. Most of the literature reviewed shows that CASA is mainly used for sound identification especially in speech recognition but CASA has also been used for sound localisation [32], [33]. As CASA uses audio cues in order to recognise the sound sources, it was considered as a possible method for acoustic feature extraction.

There are two stages to CASA techniques. The first stage is the extraction of the audio features while the extracted features are grouped in the second stage. Initially, the sound sources are filtered by filterbanks modelled after the human cochlea and divided into subbands. The features are grouped in the second stage. Groups of these features are known as streams which can be used for recognition and scene understanding[34]. In [33], sound localisation is achieved via Interaural Phase Difference (IPD) and Interaural Intensity Difference (IID) estimation. Estimated Scattering Theorem is used to analyse the estimated IPDs and IIDs depending on the subband frequency of the data [33]. IPD is used to localise sound below 1500 Hz whereas IID is used for sounds above 1500Hz as the ambiguity of the sound's direction of arrival arises for IPD when the frequency is over 1500 Hz [33].

CASA's strength lies in the way it synthesizes audio data in a way which is similar to human hearing. As such, CASA not only separates and localises the sound mixture but it also attempts to understand the audio scene presented. CASA's resemblance to the human hearing from a neurobiological perspective might also give it an edge in sound separation [34]. The main drawback of CASA is the difficulty in modelling a system resembling the human hearing with low computational complexity [34]. As this research aims to separate sound sources online with relatively low computation complexity, CASA is deemed unsuitable for the applications required in this research. However, the ideas from CASA such as the use of acoustic cues for localisation and masking can be utilised in this research. [35] takes the inspiration from CASA in the

creation of TF masks used to segregate speech sources.

2.3.4 Source Separation via Signal Sparsity

A BSS technique which is not statistically as demanding as HOS or SOS nor as computationally demanding as CASA is source separation via signal sparsity. It is a relatively new technique in the field of BSS [4]. As the name implies, this technique exploits the sparseness of the sound sources in the TF domain to achieve source separation. The sparseness of source signals or the *approximate W-disjoint orthogonality* basically means that the active source signals do not overlap too much in the TF domain. This BSS technique is capable of extracting sound sources even in underdetermined scenarios as long as the signal assumptions are satisfied. Due to the signal sparseness, it is assumed that the relative acoustic features extracted will clutter around true acoustic features [36]. As a result of this, the number of sound sources can be determined from the number of acoustic feature clusters. The acoustic features will be the peak or centers of these acoustic feature clusters. Sound separation via signal sparseness is achieved by using the source of interest's acoustic features to construct a TF mask that masks the other signals' contribution in the TF domain. The source of interest will be separated from other sources when the masked mixture is inversely transformed back from the TF domain.

Table 2.1 summarises the different source separation techniques discussed. The aim of this research is to provide an online recursive solution to separate a time varying number of speech sources so the technique chosen to extract acoustic sources has to be able to function without *a priori* knowledge of the sound sources. As the number of sound sources for the "conference room scenario" is assumed to not be known *a priori* and constantly changing, the acoustic scenarios range from overdetermined to underdetermined depending on the active speech sources at any given time.

Among all the source separation techniques reviewed, source separation via signal sparseness is deemed to be the most suitable approach to be used for the source separation component in this research as it meets most of the criteria laid out by this research. In terms of acoustic extraction, this technique is capable of extracting acoustic features by exploiting the sparseness of the source signals. Moreover, source separation via

Table 2.1: Comparison of the Different BSS Techniques

BSS technique	Number of Sources that can be Separated	Signal Assumption	Computational Complexity
HOS	$N \leq M$	Statistical independence	High
SOS	$N \leq M$	Non-stationarity	High
CASA	$N \geq M$	Human hearing inspired	Very High
BSS via signal sparsity	$N \geq M$	Signal sparsity in TF domain	Low

signal sparseness techniques are capable of extracting acoustic features in over determined, determined and even underdetermined scenarios, fulfilling the criterion of extracting acoustic features for a time varying number of speech sources. Furthermore, the computational complexity of BSS using signal sparsity is low so it fulfills the criterion of an online solution. A detailed analysis on the theory as well as the workings of source separation via signal sparseness will be given in Chapter 3. With a technique shortlisted for acoustic feature extraction for non-moving sound sources, the next problem that needs to be addressed is to extract these features when the sound sources are moving.

2.3.5 Source Separation of Moving Sound Sources using Blockwise Batch Processing

With the technique for acoustic feature extraction shortlisted, the next problem to be solved is to extract these acoustic features when the sound sources are moving. The main hurdle that makes the separation of moving sound sources more difficult than non-moving sound sources is the change in the acoustic features as the sound sources move. Blockwise batch processing is a possible method of extracting acoustic features of moving sound sources. The work presented in [37] was the first to utilize blockwise processing of a source mixture to achieve sound source separation. The work was further extended in [38] and [39]. The basic idea of these sound source separation system was to break a whole sound source signal into blocks and perform standard source separation techniques such as ICA [19] to obtain the mixing parameters of the sound source mixture. The details of source separation via ICA can be obtained in

section 2.3.1. Using the source mixing parameters extracted from the standard BSS techniques, source localisation is performed on the different sound sources with on each block of signal presented to the algorithm. The primary idea introduced by these work is that for a short enough time frame, a moving sound source can be considered as a stationary sound source within that time frame [37]. As a result of this, the length of the block or time frame chosen for analysis is extremely important. If the block is too long, the assumption of pseudo-stationarity in the sound mixture does not hold and it leads to a degradation in the source separation performance. On the other hand, if the time frame chosen is too short, there is not enough information contained within the time frame to extract the mixing parameters of the source mixture. Hence, the analysed block length should be short enough that the sound sources are pseudo-stationary yet long enough so that adequate data can be extracted from the block of signal [37].

The use of a blockwise processing method has opened up the possibility of localising and separating moving sound sources. However, these conventional techniques suffer from a common limitation: the lack of proper identity association for the separate sound sources with each frame. In [37], the separated sound sources were classified using a machine learning technique - Continuous Hidden Density Markov Models. The use of training data to associate separated speech is not practical in real life as the identity of the speakers are usually not known prior and the training of the machine adds a layer of computational complexity to the algorithm. [38] does away with the need for training data and solves the scaling and permutation issue inherent to BSS via ICA using analytical calculations for null directions. The permutation and scaling problem is solved for each frame but there is no explicit identity association between these sound sources that follows frame to frame. [39] performs data association through the combination of their proposed Number of Source Estimation Technique (NOSET) and k-means clustering [40]. The solution of data association via clustering solves the identity ambiguity of the sound sources for each time frame but there is still no reliable method of performing identity association for the separated sound sources across the frames.

There are other limitations faced by these techniques which make them an unsuitable choice to track and separate multiple time varying sound sources. [37] and [38] are techniques which only work on overdetermined or determined scenarios. The use

of NOSET [40] improves the source estimation to underdetermined scenarios but it still suffers from the problem of correctly estimating the true number of sound sources if certain sound sources dominate the TF histogram. The selection of time frequency points with dominant sound sources as a basis for source separation results in weaker sources being discarded from the mixture. [39] has another weakness in the assumption that the sound sources are near field and there is a strong direct path while the multipath components are weak. Such a scenario is usually not present in the real world environment as the speakers tracked are usually far field when they move around in a room. Although the techniques [37] [38] [39] are not incorporated into this research, the idea of pseudo-stationarity introduced by these research is adopted for the acoustic feature extraction of moving targets in the proposed algorithm. The source state dependent acoustics features, $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$, can be simplified to just a_{mn} and δ_{mn} when the sound sources are pseudo-stationary.

2.4 Sound Source Localisation

Based on the earlier reviews, a source separation technique exploiting signal sparseness processed in an online blockwise manner can extract acoustic features, $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$, from a time varying number of speech sources which are moving. However, these features alone are not enough to separate the speech sources which are moving as these features are corrupted by noise and reverberation. As shown in equation (2.8), the acoustic features, $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$, are dependent on the state, $x_{t,n}$. The change in the sound source's location will affect the acoustic features received by the microphones. Hence, the relationship between a sound source's location and its acoustic features can be utilised to identify the acoustic features associated with the sound source.

The established methods of sound source localisation can generally be categorized into three main categories: Time Difference of Arrival (TDoA) estimation methods, high-resolution spectral estimation concepts and steered response power (SRP) of beamformers [41]. The first category utilizes the delay between various combinations of microphones to determine the location of the sound source. The signature technique that represents this category is Generalised Cross-Correlation (GCC) [42].

The second category exploits the signal correlation matrix in order to achieve source localisation. An example of this technique is MUltiple SIgnal Classification (MUSIC) [43]. The last category estimates the sound location from the filtered, weighted and summed version of the signal data received at the sensors [41]. A combination of a steered beamformer with phase transform (SRP-PHAT) [41] is an example of sound localisation using beamforming.

2.4.1 Generalized Cross-Correlation (GCC)

Among the various sound source localisation technique, the GCC technique is one of the most widely used techniques. GCC is a sound localisation method proposed by Knapp and Carter in 1976 [42]. It is one of the most common techniques used for sound localisation [44]. GCC is usually used in conjunction with other weighting functions such as Maximum Likelihood (ML) [45], Smoothed Coherence Transform (SCOT) and Phase Transform (PHAT) to obtain a more accurate direction of arrival [42]. The PHAT weighting function is the magnitude of the GCC [42]. PHAT is the more commonly used weighting function as it is more robust in reverberant environments though its performance is sub-optimal in ideal conditions [46].

GCC uses a simplified version of the anechoic acoustic model with the assumption that there is only a delay between the pair of observed signals and there is no attenuation factor. The signal received by a microphone pair can be modelled as

$$\begin{aligned} y_1(t) &= s(t) + v_1(t), \\ y_2(t) &= s(t-\delta) + v_2(t), \end{aligned} \tag{2.12}$$

with $y(t)$ being the received signal, $s(t)$ the sound source and $v(t)$ the background noise. δ is the delay or TDoA between the first microphone and the second microphone. The cross correlation function $R_{\text{gcc}}(\tau)$ is used to extract the TDoA between the two microphone signals. The GCC function is described in equation 2.13

$$R_{gcc}(\delta) = \int_{-\infty}^{\infty} \Phi(\omega) \mathcal{S}_{y_1,y_2}(\omega) e^{j\omega\delta} d\omega, \tag{2.13}$$

where $R_{\text{gcc}}(\delta)$ is the cross correlation function, $\Phi(\omega)$ is the weighting function and $\mathcal{S}_{y_1,y_2}(\omega)$ is the cross spectral density of the microphone signals $y_1(t)$ and $y_2(t)$. The main weighting functions, $\Phi(\omega)$, examined in [42] are the Roth Processor [47], SCOT

and PHAT. The PHAT weighting whitens the microphone signals by giving equal emphasis across all the frequencies. PHAT weighting is usually robust in reverberant environment as it is particularly effective at reducing the degradations due to multi-path signals. The PHAT weighting can be expressed as

$$\Phi(\omega) = \frac{1}{|\mathcal{S}_{y_1,y_2}(\omega)|} \tag{2.14}$$

The TDoA between two signals can be estimated by finding the time difference $\hat{\delta}$ that maximizes the argument of the the cross correlation function

$$\hat{\Delta} = \underset{\Delta \in [-\Delta_{\max}, \Delta_{\max}]}{\arg\max} \; R_{\text{gcc}}(\Delta) \tag{2.15}$$

whereby $\Delta_{\max}$ is the maximum possible TDoA between the microphones in the pair. The direction of arrival can be estimated by utilizing the TDoA which is obtained from the features extracted. With this information,the sound source's angle of arrival can be calculated using formula 2.16 [48] [49]

$$\hat{\theta} = \sin^{-1}\left(\frac{\hat{\Delta} \times c}{d}\right) \tag{2.16}$$

where $\hat{\theta}$ refers to the estimated DoA, $\hat{\Delta}$ refers to the estimated TDoA, c refers to the speed of sound and d refers to the distance between the two microphones. This theory only holds true if the sound sources are located in the far field as this theory assumes the sound waves arriving at both the sensors are planar waves. Refer to figure 2.2 for the relationship between the DoA and TDoA.

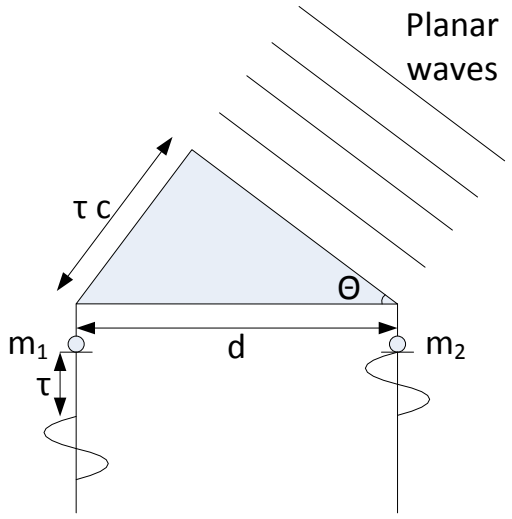


Figure 2.2: Relationship between DoA and TDoA

Generally, GCC-PHAT is usually applied in real life to determine speaker position in a room [50] [51]. In [50], a modified version of GCC-PHAT is used to construct the acoustic map of the room in order to localise two speakers. In [51], GCC-PHAT is combined with High and Low Band energy Ratios (HLBR) for speaker localisation. The combination retains the simplicity of HLBR while exploiting the robustness of GCC-PHAT. GCC has undergone many modifications over the years but the basic idea remains the same - the TDoA can be obtained from the cross power spectral density spectral function of the two input signals. GCC-PHAT has been shown to be robust in reverberant environments [46]. The main drawback to the original GCC-PHAT is its incapability in localising more than one sound source without further modifications [44] [50]. Moreover, the solution of the GCC is uniquely complex due to equation (2.15). As this research intends to localise, track and separate multiple time varying number of sound sources, a conventional GCC-PHAT might not suit the purposes of this research despite its capability in determining the TDoA of a sound source which can be used to calculate the delay, δ, of a source signal.

2.4.2 MUSIC

Apart from GCC, MUltiple SIgnal Classification (MUSIC) [43] is another method of determining the DoA, θ, of a sound source. Using the relationship between DoA and TDoA as shown in equation (2.16), the delay, δ, of a source signal can be determined once the DoA is estimated.

MUSIC is a subspace parameter estimation techniques that can be used to determine the number of sound sources and the DOA of sound sources. MUSIC relies on eigenvalue decomposition to determine the number of sound sources and the DOA of a particular signal [52, p.1158][53]. MUSIC decomposes the matrix of the received signal into two orthogonal matrices - the signal subspace and the noise subspace. As MUSIC was developed for narrowband signals, the received signals can be modelled as

$$\begin{aligned} y_1(\omega) &= s(\omega) + v_1(\omega), \\ y_2(\omega) &= s(\omega)\, e^{-j\omega\delta} + v_2(\omega) \end{aligned} \tag{2.17}$$

in the frequency domain. The data model used by MUSIC is similar to the general

BSS model in equation (2.1) with the addition of a noise and it is expressed in vector form

$$\mathbf{y}(\omega) = \mathbf{A}_{\text{steering}}(\theta)\mathbf{s}(\omega) + \mathbf{v}(\omega) \tag{2.18}$$

$$\begin{bmatrix} y_1(\omega) \\ y_2(\omega) \\ y_3(\omega) \\ \vdots \\ y_M(\omega) \end{bmatrix} = \begin{bmatrix} a_1(\theta_1) & a_1(\theta_2) & a_1(\theta_3) & \cdots & a_1(\theta_N) \\ a_2(\theta_1) & a_2(\theta_2) & a_2\theta_3) & \cdots & a_2(\theta_N) \\ a_3(\theta_1) & a_3(\theta_2) & a_3(\theta_3) & \cdots & a_3(\theta_N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_M(\theta_1) & a_M(\theta_2) & a_M(\theta_3) & \cdots & a_M(\theta_N) \end{bmatrix} \begin{bmatrix} s_1(\omega) \\ s_2(\omega) \\ s_3(\omega) \\ \vdots \\ s_N(\omega) \end{bmatrix} + \begin{bmatrix} v_1(\omega) \\ v_2(\omega) \\ v_3(\omega) \\ \vdots \\ v_N(\omega) \end{bmatrix}$$

where $\mathbf{y}$ is a $M \times 1$ vector of received signals, $\mathbf{A}_{\text{steering}}$ is a $M \times N$ matrix of steering vectors, $\mathbf{s}$ is a $N \times 1$ vector of source signals, and $\mathbf{v}$ is a $M \times 1$ vector of additive noise.

Based on the assumption that the elements in $\mathbf{A}_{\text{steering}}$ which is in the signal subspace is orthogonal to the noise subspace, eigenvalue decomposition is performed on the signals. MUSIC works by plotting the pseudospectrum of the signal and searching through all the angles

$$P_{\text{MUSIC}}(\theta) = \frac{1}{a^H(\theta)\mu_n} \tag{2.19}$$

where $a(\theta)$ is the steering vector, $[\cdot]^H$ is the Hermitian transpose and μ_n is the eigenvalue. The signals' DOA can be estimated by searching for the largest peaks as the denominator of the equation (2.19) will be zero when θ is in the signal direction. By rearranging equation (2.16), the delta can be calculated from the estimated theta with equation (2.20).

$$\hat{\delta} = \frac{d \sin \hat{\theta}}{c} \tag{2.20}$$

The estimated delay, $\hat{\delta}$ is one of the acoustic features of a source with the other being the attenuation. Apart from sound localisation, MUSIC can also be used to estimate the delay of a sound source. Root MUSIC [54] is a variation of spectral MUSIC that estimates the DOA of the signals by solving the roots of a polynomial instead of searching the pseudospectrum for peaks [55]. Root MUSIC can be applied when the sensor array is a uniform linear array.

Although spectral MUSIC and root MUSIC have been shown to be capable in estimating the DoA of signals for any sensor array geometry, it is not a suitable sound source localisation technique for this research. This is mainly due to the fact that the MUSIC algorithm is computationally expensive to run. In order to determine the

source locations, it needs to search the whole pseudo-spectrum. With a more complex signal model that is required in a more realistic acoustic environment, computational complexity of this algorithm will increase. Furthermore, MUSIC was initially developed for narrowband signals and computational complexity will increase if the technique is to be extended to wideband signals in [56] and [57]. The MUSIC algorithm is not very robust as it is sensitive to source and sensor modeling errors and it requires calibration in order to perform well [58]. Another limitation of MUSIC is its incapability of estimating the DOA in underdetermined scenarios as it also makes the assumption that the number of sources is less than the number of microphones. The focus of this research is in estimating a time varying number of sound sources online, it is impossible to determine the number of sensors required as the number of sound sources vary with time and there is no prior knowledge on the number of sound sources to be tracked. It is due to these various factors that MUSIC is not considered as a source localisation technique for this research.

2.4.3 Steered Beamformer with Phase Transform (SRP-PHAT)

Apart from MUSIC, beamforming is another narrowband technique used for sound localisation purposes. The beamformer has been around since the 70s as a method to determine signal's DoA [59]. The adaptive beamformer which is more robust than the conventional beamformer was introduced in the 80s [60] [61]. Compared to the conventional beamformer which was susceptible to interference signals, the adaptive beamformer is more robust as it is capable of nulling these jammer signals. The beamformer has been applied in various areas such as radar, sonar, communication and astrophysical exploration [59]. The conventional beamformer is also known as the delay and sum beamformer [55]. The propagation delays in the arrival of the source signal at each of the sensors in the array is compensated using time-shifts. The signals are then aligned and summed to form a single output signal [41]. Filters are applied to the conventional delay and sum beamformer in order to enhance the primary beam and the different filter and sum beamformers are differentiated based on the type of filters used [41]. An example of a beamformer is illustrated in figure 2.3.

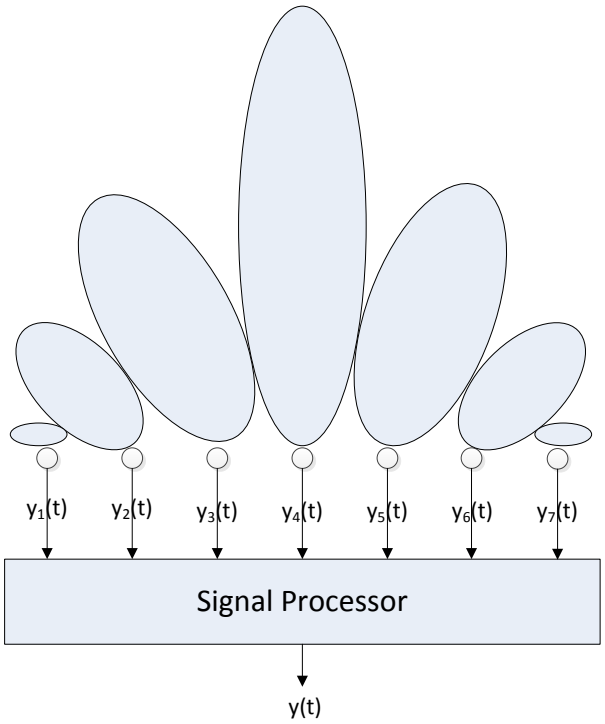


Figure 2.3: An example of a beamformer

For speech type applications, beamforming was mostly used for voice capture rather than for sound source localisation due to the higher efficiency of other sound localisation methods. An SRP beamformer relies on the spectral content of a sound source in order to perform source localisation. However, this is impractical as a priori knowledge of the source signal and the background noise's spectral content is required to derive the optimal solution [41]. This is further compounded by the presence of reverberation in the received signal as it is difficult to estimate the background noise accurately. To this end, the work presented in [41] addressed the problem by introducing the SRP-PHAT beamformer - a combination of the steered beamformer with phase transform.

The SRP-PHAT beamformer proposed in [41] combines the advantages of steered beamformer with those of PHAT. The steered beamformer is much more resistant towards the effects of environmental conditions and requires shorter analysis intervals suffers performance deterioration in the presence of multi-path signals in a room as it relies heavily on the prior knowledge of the signal and channel content. The PHAT weighting function which is resistant towards low to moderate reverberant environment is able to mitigate the effect of multipath on the performance of the steered beamformer

while the beamformer complements the PHAT weighting by decreasing its sensitivity towards the effects of environmental conditions.

For an N-element SRP-PHAT beamformer, the output in the frequency domain can be expressed as [41]

$$\mathcal{P}(\mathfrak{p}) = \sum_{l=1}^{N}\sum_{k=1}^{N}\int_{-\infty}^{\infty}\Phi_{lk}(\omega)\mathcal{S}_{y_l,y_k}(\omega)e^{j\omega\delta}d\omega \tag{2.21}$$

with $\Phi_{lk}(\omega)$ being the PHAT weighting function, $\mathcal{S}_{y_l,y_k}(\omega)$ referring to the cross spectral density of the received signals, y_l and y_k while δ refers to the TDoA between the l^{th} and the k^{th} microphones. The PHAT weighting used in SRP-PHAT is similar to the one used in equation 2.22. Its multi-channel counterpart can be expressed as

$$\Phi_{lk}(\omega) = \frac{1}{|\mathcal{S}_{y_l,y_k}(\omega)|} \tag{2.22}$$

In the context of filter-and-sum beamformers, the PHAT weighting function is the same as using individual filter channels

$$G_n(\omega) = \frac{1}{|Y_n(\omega)|} \tag{2.23}$$

with $G_n(\omega)$ referring to the filter and $Y_n(\omega)$ referring to the STFT of the n^{th} microphone signal. Similar to the conventional SRP beamformer, the location of a sound source is estimated by scanning the power across all the potential regions where the sound source might be located and maximizing the power

$$\hat{\mathfrak{p}}_s = \arg\max_{\mathfrak{p}}\mathcal{P}(\mathfrak{p}) \tag{2.24}$$

with $\hat{\mathfrak{p}}_s$ referring to the estimated sound source position.

The advantage offered by SRP-PHAT beamformer is the capability to localise sound sources in a reverberant environment as it significantly de-emphasises the extraneous peaks and boosting the peak of the true sound source signal. With the combination of the PHAT weighting function, the SRP-PHAT beamformer is more resistant to the effects of noise and reverberation. The main drawback of the SRP-PHAT is the additional computational cost involved [41] due to its large search space. Moreover the large search space of the SRP-PHAT also produces many extremas [62]. ROOT-SRP-PHAT [63] is a technique that reduces the computational load of the SRP-PHAT through the use of root solving. The experiment carried out in [63] shows ROOT-SRP-PHAT to be much less computationally intensive despite being almost as robust

as SRP-PHAT. Despite having a robustness close to SRP-PHAT at less computational complexity, the ROOT-SRP-PHAT was not used as it is less reliable when the room reverberation is high.

Based on the investigation carried out on the different sound source localisation techniques, the techniques would not be directly used in this research mainly because the source localisation techniques require the number of microphones to be more than the number of sound sources and the number of sound sources are not known *a priori* in the context of this research. However, the main idea espoused by theses techniques, namely the relationship between a sound source's location and its delay can be exploited to estimate a sound source's position, will be adopted in this research. As the delay of a sound source, δ, can be estimated using source separation techniques, the DoA, θ, of sound source can be determined from the estimated delay using equation (2.16). As the aim of this research is to estimate the sound sources' states rather than just their DoAs, more sophisticated techniques which are able to further exploit the relationship between the delay of a sound source and its location are required. Furthermore, the sound sources need to be tracked across time so traditional localisation techniques such as GCC, MUSIC and SRP-PHAT which assume the sound sources to remain in the same position are ill suited of the task of tracking.

2.5 Bayesian based Sound Source Tracking and Separation

In order to track the states of the moving sound sources, Bayesian sound source tracking techniques are investigated. Bayesian sound source tracking makes use of state-space estimation techniques such as the Sequential Monte Carlo (SMC) to achieve sound source tracking. In state-space estimation techniques, the sound source's location and velocity are modelled as a state vector and the transition of this state is predicted and updated with observations made. The prediction of a sound source's trajectory based on its dynamic model produces a more accurate estimate of the sound source at a given time. Most state space estimation techniques usually combine a sound localisation technique with a tracking technique by using the information extracted using the source localisation technique as observations for the tracking technique. Hence,

source separation techniques and tracking techniques can complement each other's functions as one is able to provide the observations required and the other is able to utilise these information to track the movement of the sound sources. Examples of state estimation techniques are sound source tracking with Particle Filters (PF) [64], sound source tracking with Random Finite Set (RFS) [65] as well as sound source tracking and separation using Multiple Acoustic Source Tracking (MAST) [66].

2.5.1 Sound Source Tracking using Particle Filter (PF)

The idea of using a Particle Filter (PF) to solve the sound source tracking problem was proposed by Ward and Lehmann in [64]. It was later updated with a Voice Activity Detector [67] in order to be more efficient in the information selection for TDoA extraction. The work by Ward and Lehmann [64] has spurred further research in the use of a PF in sound source tracking [68][69][70]. The motivation of using a state space estimation for sound source tracking was the dynamic nature of a sound's direct path. Traditional sound source localisation techniques such as GCC [42], MUSIC [43] and beamforming tend to estimate a sound source's localisation at just the given time frame and such estimation will be affected by noise and reverberation. Given the fact that there is a spatio-temporal correlation between the sound source's direct path whereas the contribution by noise and reverberation is spurious in nature, sound source tracking with a PF is able to mitigate the effects of noise and reverberation on the sound source's localisation accuracy [64].

Sound source tracking using PF algorithm is executed as a Bayesian recursion [71]. The general framework of sound source tracking using PF is an estimation of a sound source's TDoA, predicting its location based on the defined dynamics model and updating the likelihood of the prediction based on the received observation and finally estimating the location of the sound source based on the weighted sum of the particles' positions [72].

A target at time k can be represented by its single target state, x_k, which contains the Cartesian Coordinates and their respective velocities

$$x_k = [\mathfrak{p}_{k,\mathrm{x_{coord}}}\ \mathfrak{p}_{k,\mathrm{y_{coord}}}\ \dot{\mathfrak{p}}_{k,\mathrm{x_{coord}}}\ \dot{\mathfrak{p}}_{k,\mathrm{x_{coord}}}]^T. \tag{2.25}$$

The single target state is assumed to follow a Markov process in the state space, $\mathcal{X} \subseteq$

$\mathbb{R}^{n_x}$, and this Markov process can be described by the transition density also known as the dynamics model, $f_{k|k-1}(x_k|x_{k-1})$ [73, p.76]. Different dynamics models were examined and compared in [74]. Based on the comparative study, the *Langevin* model was deemed to be the most appropriate dynamics model to be used. The *Langevin* model assumes each Cartesian Coordinate to be independent [64].

The Markov process undergone by the state can be partially observed from the measurements, z_k, in the observation space, $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$. The likelihood function, $g(z_k|x_k)$, describes the probability that a measurement, z_k, is generated by the state, x_k. In PF, the measurements, z_k, are usually generated by source localisation techniques. Different source localisation methods can be used in conjunction with a particle filter and these source localisation techniques can be divided into two categories: TDoA estimation techniques and direct localisation techniques [75]. An example of TDoA estimation is GCC while an example of a direct localisation technique is beamforming [75]. In order to relate the observed measurement with the state estimate, equation 2.26 is used [76]

$$\hat{\Delta}_k = \frac{|\hat{\mathfrak{p}}_{s,k} - \mathfrak{p}_{\mathrm{mic},m}| - |\hat{\mathfrak{p}}_{\mathbf{s},\mathbf{k}} - \mathfrak{p}_{\mathrm{mic},\mathbf{n}}|}{c} \tag{2.26}$$

with $\hat{\Delta}_k$ referring to the estimated TDoA between the microphone pair at time k, $\hat{\mathfrak{p}}_{s,k}$ being the estimated location vector of the sound source which contains the XY coordinates, $\mathfrak{p}_{\mathrm{x_{coord}},k}$ and $\mathfrak{p}_{\mathrm{y_{coord}},k}$, $\mathfrak{p}_{\mathrm{mic},m}$ and $\mathfrak{p}_{\mathrm{mic},n}$ referring to the microphone pair and c is the speed of sound.

The likelihood to be used in PF can either be a Gaussian likelihood or pseudo-likelihood [64]. A Gaussian likelihood is formed through the assumption that the true source is corrupted by the addition of a Gaussian noise so the likelihood is a Gaussian function with the true location of a particular sound source as a mean. On the other hand, a pseudo-likelihood is formed by using the continuous function of the localisation as function of density to measure the likelihood. A lower bound is introduced in the pseudo-likelihood to account for cases whereby there are no peaks which correspond to the true sound source's location. Both the Gaussian likelihood and the pseudo-likelihood can be paired with the two conventional types of source localisation techniques. A comparison of the different combinations is reviewed in [64]. The results show that a combination of a direct localisation method paired with a pseudo-likelihood performed best in the 100 simulation runs performed although the authors

do concede that further verification of the performance is required with real recordings [64].

The evolution of the single target state can be described by the posterior density, $p_k(x_k|z_{1:k})$. Based on the measurements up to time k, the posterior density, $p_k(x_k|z_{1:k})$, can be calculated from the the Bayes recursion

$$p_{k|k-1}(x_k|z_{1:k-1}) = \int f_{k|k-1}(x_k|x)p_{k-1}(x|z_{1:k-1})dx \tag{2.27}$$

$$p_k(x_k|z_{1:k}) = \frac{g_k(z_k|x_k)p_{k|k-1}(x_k|z_{1:k-1})}{\int g_k(z_k|x)p_{k|k-1}(x|z_{1:k-1})dx} \tag{2.28}$$

which relies on the information provided by the dynamics model and also the likelihood function. Pseudocode 1 illustrates the recursive stages of sound source tracking using PF.

Algorithm 1 Pseudocode for Sound Source Tracking using Particle Filter

Data Initialization

initialise the algorithm by forming a set of particles $\{\mathscr{X}_0^{(l)}\}_{l=1}^{L}$ with an initial uniformly distributed weight, $\{w_0^{(l)}\}_{l=1}^{L} = \frac{1}{L}$, with L being the number of particles

Filtering

while Frame of data, k, examined is not the last frame,K **do**

Resample the particles from the previous frame,$\{\mathscr{X}_{k-1}^{(l)}\}_{l=1}^{L}$, according to their weights in the previous frame to form a new set of particles, $\{\tilde{\mathscr{X}}_{k-1}^{(l)}\}_{l=1}^{L}$

Predict the set of particles, $\{\mathscr{X}_k^{(l)}\}_{l=1}^{L}$, by propagating the particles, $\{\tilde{\mathscr{X}}_{k-1}^{(l)}\}_{l=1}^{L}$, through the chosen dynamics model, $f(x_k|x_{k-1})$

Transform the observed raw data, δ, by the localisation function into a localisation measurements, z

Using the likelihood function, $g_k(z|x)$, calculate the likelihood based on the observed measurements, z

Weight the particles according to the new likelihood, $w_k^{(l)} = g_k(z|x)$, and normalize the particle weights to 1, $\sum_l w_k^{(l)} = 1$.

Store the set of particles and their respective weights, $\{\mathscr{X}_k^{(l)}, w_k^{(l)}\}_{l=1}^{L}$

State Estimation

Estimate the sound source's location based on the weighted sum of the particle positions

end while

Despite the improvements introduced by [64] in the use of PF to track sound source movements, there are limitations to this technique. One of the limitations of this technique is the number of sound sources that can be tracked. [64] and [67] are only capable of tracking one sound source. This limitation on the number of sound sources which can be tracked needs to be addressed as there are usually more than one sound source of interest in real world environments. As a result of this, the capability of the sound source tracker with PF is generalised in [77] through the use of Random Finite Sets (RFS).

2.5.2 Sound Source Tracking using Random Finite Sets (RFS)

In a real world acoustic scenario, the number of speakers in a room will be unknown and time varying as speakers will randomly appear and disappear. A joint estimation on the number of sound sources as well as the sound sources' location is required for the purpose of tracking multiple speakers in a room. As a result of this, the work presented in [64] which only tracks a single target has its limitations in a real world application. There were earlier attempts to use Kalman filters to track multiple speakers as shown in the work presented in [78] and [79]. However, in order to properly track multiple time varying number of sound sources in a room, a multi-target tracking technique would be required.

The multi-target tracking research field is well established with three main methods of tracking multiple targets - Multiple Hypothesis Tracker (MHT) [80], Joint Probabilistic Data Association (JPDA) [81] and Random Finite Sets (RFS) [12]. In MHT, the data association hypotheses is propagated in time to achieve multi-target tracking. Multi-target tracking using JPDA is achieved through the propagation of observations based on their association probabilities. Both MHT and JPDA are computationally expensive to run in an online environment. Multi-target tracking with MHT is computationally expensive because it enumerates all the possible hypotheses and the total number of possible hypotheses increases exponentially with time [82]. In JPDA, the marginal joint association probabilities is the sum of all the joint association probabilities. Thus, JPDA is also computationally expensive to run when the number of targets tracked increases as the calculation of the joint association probabilities scales exponentially with the number of targets and the number of measurements [83]. Moreover, JPDA also does not explicitly handle target birth and death [84] so it is not a suitable technique to detect speech sources which may appear or disappear with time.

In order to extend the single target tracking problem to a multi-target problem, the most suitable method would be to represent the number of states and their values as random finite sets (RFS). An RFS is a finite-set-valued random variable which is characterised by its probability distribution [13]. The motivation of using finite sets stems from the notion of estimation error [85] [73, p.76]. In estimation theory, the output of an estimation has little meaning if there is no meaningful way of interpreting the estimation error [86]. Extending a single state to a multi-state through the use of a

vector has fundamental inconsistencies in terms of estimation error. The use of a vector to represent a multi-state has problems accounting for the estimation errors in terms of data association, accuracy of the estimated states, as well as the cardinality of targets. On the other hand, representing the multi-target states as finite sets is consistent with the notion of estimation error as it is able to represent all the possible combinations of the multi-target state and the miss distance between sets which accounts for the accuracy of the estimated states and the cardinality of the estimate [73, p.76]. The problem of multi-target tracking using RFS can be solved using the Finite Set Statistics (FISST) which was proposed by Mahler in [12].

FISST was developed based on the notion of integration and density, which is consistent with point process theory [73, p.76].

Multi-target System Model

In single target tracking such as the technique used in [64], a target is usually represented by its single target state, x_k, which consists of the position and velocity of the moving target at time k. The single target transitions through the state space, $\mathcal{X} \subseteq \mathbb{R}^{n_x}$, and its movement can be observed in the observation space, $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$. When the number of the targets is $N(k)$, the target states $x_{k,1}, ..., x_{k,N(k)}$ and the measurements as well as clutter generated by these targets, $z_{k,1}, ..., z_{k,M(k)}$ can be represented as [13] [12][87]

$$X_k = \{x_{k,1}, ..., x_{k,N(k)}\} \in \mathcal{F}(\mathcal{X}), \tag{2.29}$$

$$Z_k = \{z_{k,1}, ..., z_{k,M(k)}\} \in \mathcal{F}(\mathcal{Z}). \tag{2.30}$$

In RFS, $\mathcal{F}(\mathcal{X})$ represents the space of the finite subsets of $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ while $\mathcal{F}(\mathcal{Z})$ represents the space of the finite subsets of $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$.

At time k, a target within the previous multi-target state, $x_{k-1} \in X_{k-1}$ may cease to exist with the probability, $1 - p_{S,k}(x_{k-1})$ or it may survive with probability $p_{S,k}(x_{k-1})$ and transition to a new state according to the transition density, $f_{k|k-1}(x_k|x_{k-1})$. These single targets are states within the multi-target set. Hence, the dynamics of a multi target state can be described by the union of a survival RFS at time k, $S_{k|k-1}(x_k - 1)$

and the multi target birth model, Γ_k, as shown in the following expression

$$X_k = \left[\bigcup_{x_{k-1} \in X_{k-1}} S_{k|k-1}(x_{k-1}) \right] \cup \Gamma_k \tag{2.31}$$

With the union of the RFS survival model and the RFS birth model, the multi-target transition density takes the target birth, target death and the target motion into consideration.

During the measurement period, a target, $x_k \in X_k$, at given time, k, can be missed by the probability $1 - p_{D,k}(x_k)$ or detected with the probability $p_{D,k}(x_k)$. If the target is detected, there is a likelihood, $g_k(z|x_k)$, it will generate an observation, z_k. The measurement generated by the target at time k can be described by the RFS, $\Theta_k(x)$. Apart from the actual measurements generated by the target, the sensors might receive clutter or false alarms, K_k with the intensity function, $\kappa_k(\cdot)$. The multi-target measurement, Z_k, generated by the multi-target state, X_k, can be expressed as a union between the measurement's RFS and the clutter's RFS

$$Z_k = \left[\bigcup_{x \in X_k} \Theta_k(x) \right] \cup K_k \tag{2.32}$$

Detection uncertainty, target measurement and clutter are taken into consideration with the union of the measurement RFS and the clutter RFS.

Bayesian Multi-target filtering

In instances where a single target is tracked using a Bayesian framework, the state of the target is propagated through a Bayes recursion. When the single target tracking scenario is extended to a scenario involving the tracking of multiple moving targets, the *multi-target posterior density* at time k, $\pi_k(\cdot|Z_{1:k})$ needs to be propagated instead of the probability density of just a single state. $\pi_k(X_k|Z_{1:k})$ is propagated according to [13] [12] [87]

$$\pi_{k|k-1}(X_k|Z_{1:k-1}) = \int f_{k|k-1}(X_k|X)\pi_{k-1}(X|Z_{1:k-1})\delta X \tag{2.33}$$

$$\pi_k(X_k|Z_{1:k}) = \frac{g_k(Z_k|X_k)\pi_{k|k-1}(X_k|Z_{1:k-1})}{\int g_k(Z_k|X)\pi_{k|k-1}(X|Z_{1:k-1})\delta X} \tag{2.34}$$

where $f_{k|k-1}(\cdot|\cdot)$ is the multi-target transition density, $g_k(\cdot|\cdot)$ is the multi-target likelihood and the integral with reference to δX, is a *set integral* for any function f :

$\mathcal{F}(\mathcal{X}) \rightarrow \mathbb{R}$ which is defined by:

$$\int f(X)\delta X = \sum_{n=0}^{\infty} \frac{1}{n!} \int f(x_1, \cdots, x_n) d(x_1, \cdots, x_n). \tag{2.35}$$

This multi-target transition density captures the target births, deaths and motion while the detection and the clutter is captured by the multi-target likelihood. The full multi-target Bayes recursion is intractable due the inherent combinatorial nature of multi-target densities and the multiple intergrations on the multi-target state and observation spaces [12]. Hence, an approximation to the full multi-target Bayes recursion is required.

Multiple Sound Sources Tracking using PHD and CPHD

In order to overcome the limitation of [64], the work in [65] and [77] proposes a generalised target tracking approach to track multiple sound sources through the use of RFS in the form of the Probability Hypothesis Density (PHD) filter [12]. The PHD propagates the multi-target first order moment statistics rather than the actual multi-target Bayes posterior [12]. [88] further extends the capabilities of [77] by using Cardinalized Probability Hypothesis Density (CPHD) filter [89], a more advanced version of RFS technique to track multiple speakers. All the speakers which are tracked are considered a set of the single state while the observations that was generated by the speakers are treated as a set of observations [65].

The cardinality or the number of elements in RFS representing the number of targets is approximated as a Poisson distribution in the PHD filter. The Poisson distribution is described by its mean which is equal to its variance. Hence, there is a high variance in the cardinality estimated by the PHD filter when the expected number of targets is high [90]. As a result of this, the number of estimated sound sources in [65] and [77] will suffer from a high cardinality variance when the number of targets is high. A Poisson distribution is a good representation of clutter but not necessarily a good representation of the cardinality of targets. The use of CPHD to track multiple sound sources in [88] will alleviate the cardinality inaccuracy. The CPHD is a generalized version of the PHD filter which jointly propagates the intensity function as well as the cardinality distribution [90]. As the movement of the speakers in a room are usually random and nonlinear, the CPHD implementation used in [88] is the nonlinear

CPHD using unscented transforms [90].

The dynamic model used by [77] and [88] to propagate the target states is the *Langevin* model. For the measurement model, both techniques use the TDoA between microphone pairs as observed measurements. The sound source localisation technique used to obtain the multi-targets' TDoAs by both techniques is the classical GCC-PHAT technique (an explanation on the GCC-PHAT can be found in Section 2.4.1). The TDoAs obtained from the microphone pairs are not directly used for localisation, rather these TDoA estimates are fed into the PHD or CPHD filters as sets of measurements. The relationship between the TDoA measurements and the location of the sound source is given by equation (2.26). Equation (2.26) is solvable as long as the positions of the microphones used in the array are known. The likelihood model used by both techniques is a normal Gaussian distribution with the observed TDoA as the mean of the Gaussian and the observation noise as the variance. As multiple sensors are used, data fusion of the likelihood is a product of the individual likelihood for the sensor pairs.

Overall, both the techniques shown in [77] and [88] have shown satisfactory results in tracking multiple time varying sound sources. Despite this, the tracking performance of multiple sound sources can be improved as the PHD and CPHD filters were initially proposed to handle more linear scenarios whereas the dynamics model and the observation model of sound sources are nonlinear. Moreover, the use of the PHD filter is computationally expensive as it requires clustering to extract state estimates from particle clusters. The limitations of the PHD and CPHD are addressed with the Cardinality Balanced Multi-target Multi-Bernoulli (CBMeMBer) [13] filter. Unlike the PHD or CPHD filter which propagates the intensity function, the CBMeMBer filter approximates the multi-target posterior density as a multi-Bernoulli and propagates it [13]. The SMC implementation of the CBMeMBer is more suited for the task of tracking targets with non-linear dynamics and speech sources usually move in a non-linear manner. As a result of this, the RFS technique that this research focuses on is the CBMeMBer technique. Further justifications and explanations on the CBMeMBer technique can be found in Chapter 5.

2.5.3 Sound Source Tracking and Separation

The techniques discussed in Section 2.5.2 only tracks multiple speech sources but do not separate these speech sources. There have been earlier attempts to localise and separate moving sound source mixtures using purely acoustic technique [66] as well as others which adopts a multimodal approach such as [91] [92] and [93].

Sound Source Tracking and Separation using Multiple Acoustic Source Tracking (MAST)

A more recent attempt to track and separate moving sound sources purely from an acoustic standpoint is the work by Pertila in [66] which uses Multiple Acoustic Source Tracking (MAST).

The MAST technique used in [66] utilises the spatio temporal nature of speech and combines it with Bayesian tracking technique to perform sound source tracking and separation. The Bayesian tracking technique used in [66] is an independently developed technique for multiple sound sources tracking [94]. MAST uses multiple particle filters to track the multiple sound sources. A likelihood ratio based on the hypothesis, $\mathcal{H}_0$ - "No source is present", and $\mathcal{H}_1$ - "Source is present" is used to determine the presence of a sound source. The likelihood ratio is expressed as

$$\lambda(\tau) = \frac{p(\mathbf{Z}_\tau|\mathcal{H}_1)}{p(\mathbf{Z}_\tau|\mathcal{H}_0)} \tag{2.36}$$

A sound source is considered present if the likelihood is more than a predetermined threshold, λ_T. MAST uses an Attack and Decay filter is applied to the likelihood ratio in order detect new sound sources faster while keeping existing sound sources for longer at the same time. Sound source separation is performed after the sound sources have been localised and tracked. The spatial filter used to separate the acoustic features of each sound source for all the frequency bands, is similar to the one used in [95].

The work presented in [66] has been shown to be a viable method in tracking and separating sound sources but there are still several limitations to it. The first limitation of this solution is the dimension of the sound sources tracked. Due to the likelihood model and the data association method chosen, only the DoA of the sound sources are estimated and tracked. The data association used is also of concern as the authors did concede that the error of target swapping and other mismatch errors during tracking

were not evaluated in [66]. Gating is not technically a data association technique as it does not establish the identity between the estimated parameter. Gating is a technique to reduce the candidate of observed measurements based on the distance from the predicted target's position [96]. Thus, the use of gating for the purposes of data association rely on the assumption that the targets do not exist close together at the same time or else there might be a mismatch in the targets' identities. Apart from that, the birth and death model chosen to be used in [66] does not represent a smooth transition as the presence and death still depends on a preset threshold of the likelihood ratio. The likelihood ratio is not consistent with the formal definition of a Bayesian probability as it uses the hypothesis as a state.

Multimodal Sound Source Tracking and Separation

In order to localise and separate moving sound source mixtures, there are research such as [91] [92] and [93] which adopted the multimodal approach. The technique introduced in [91] can be considered a pioneering technique in multimodal sound source tracking. These methods are not purely "acoustic" tracking as they rely on additional information such as image processing in order to obtain the required information to track the sound sources. In order to extract audio information, these methods generally use classical sound localisation approaches with a combination of Bayesian sound source tracking techniques.

In [91], the multimodal approach was used to track speaker movements but the movements were limited. The solution was used to track the head movement of the speaker. Sound source localisation was achieved using GCC (Section 2.4.1 contains the details on source localisation using GCC). A PF is used to fuse the audio and visual information so the spurious peaks due to reverberation do not pose significant effects on the accuracy of the algorithm. The TDoA information obtained is used to initialise the state of the source and also helps to recover from the loss of lock [91]. A more advanced version that tracks that movements of multiple speakers in the room is developed in [92]. The sound source localisation technique is a two step approach with a sector based localisation to determine the active sector and then a point based localisation to determine the speaker's position in the sector. The sector localisation technique used is SAM-SPARSE-MEAN [92] while the point based localisation technique used

is SRP-PHAT [41] (details of the SRP-PHAT can be found in Section 2.4.3). The tracking and identity association between the different speech sources is carried using a multimodal Bayesian Information Cue which fuses the acoustic cue and the location cue of the speech sources to determine their identities. Visual cue is used to further enhance the data association of the separated speech sources. It is especially useful when recovering the tracks from the silence periods in speech sources. Unlike [91] and [92], source localisation and tracking in [93] is performed using the visual data rather than the audio data. The tracking method employed is the PF. State information of the sources such as their location and velocity is obtained through image processing. The trajectory information is then used to aid the ICA method in separating the moving sound sources.

The multimodal methods have been shown to be capable of localising, tracking and separating multiple speech sources but these methods are not without their limitations. The main problem with these methods is the additional processing required such as speech recognition or image recognition in order to identify the separated sound sources. These would mean additional computational complexity in order to identify the separated sound sources and they will be only suited for offline processing. A more specific example of this would be the Bayesian Information Criterion employed in [92] to cluster the speech sources. In order for the acoustic cue to be used for data association, the speech signal analysed has be active for a minimum duration of time. The proposed scheme of using location and acoustic cues are close to the one that only uses the location cue [92]. Hence, the acoustic tracking and separation method presented in [92] is not suitable for an online solution for speech source separation.

2.6 Proposed Solution

Based on the review of the literature in the field of source separation, source localisation and tracking, a possible solution is to model the number and the state of the acoustic targets which can vary over time as RFS and solve for these parameters using a principled RFS target tracking algorithm based on the acoustic information extracted using a BSS technique. As discussed in Section 2.1, the system does not have even have *a priori* knowledge of the sound sources so future knowledge of the sound sources

is even more unlikely. Thus, the solution also needs to be online as the estimated number and state of the acoustic targets are based on the acoustic features available to the system up to current time and no future data is available to the system. In the "conference room problem", the solution needs to be able to estimate the correct acoustic features and location of the speakers as they are speaking. The number and location of the speakers are recursively estimated as the sequence of speech data is received. According to [10], the recursive solution can be considered an online solution to the Bayesian learning problem as the information on the parameters is updated through the use of new pieces of incoming information. Figure 2.4 illustrates the workings and the the online nature of the proposed solution Both the sound source separation technique and the RFS target tracking technique complement each other in achieving the objectives of tracking and separating a time varying number of moving sound sources in real time. Acoustics features of the sound source mixture are to be extracted and used as observed measurements for the target tracking algorithm for the current time instance. By exploiting the relationship between the sound sources' acoustic features and their locations, acoustic feature such as the relative delay can be used by the target tracking algorithm to estimate the number of targets as well the state of these targets. The spatio-temporal relationship of the sound sources can be further exploited to determine a more accurate set of acoustic features from the output of the target tracking algorithm. The TF mask construction relies on the estimated number of targets and the set of filtered acoustic features. Sound source separation is achieved by applying the TF masks on the sound source mixture.

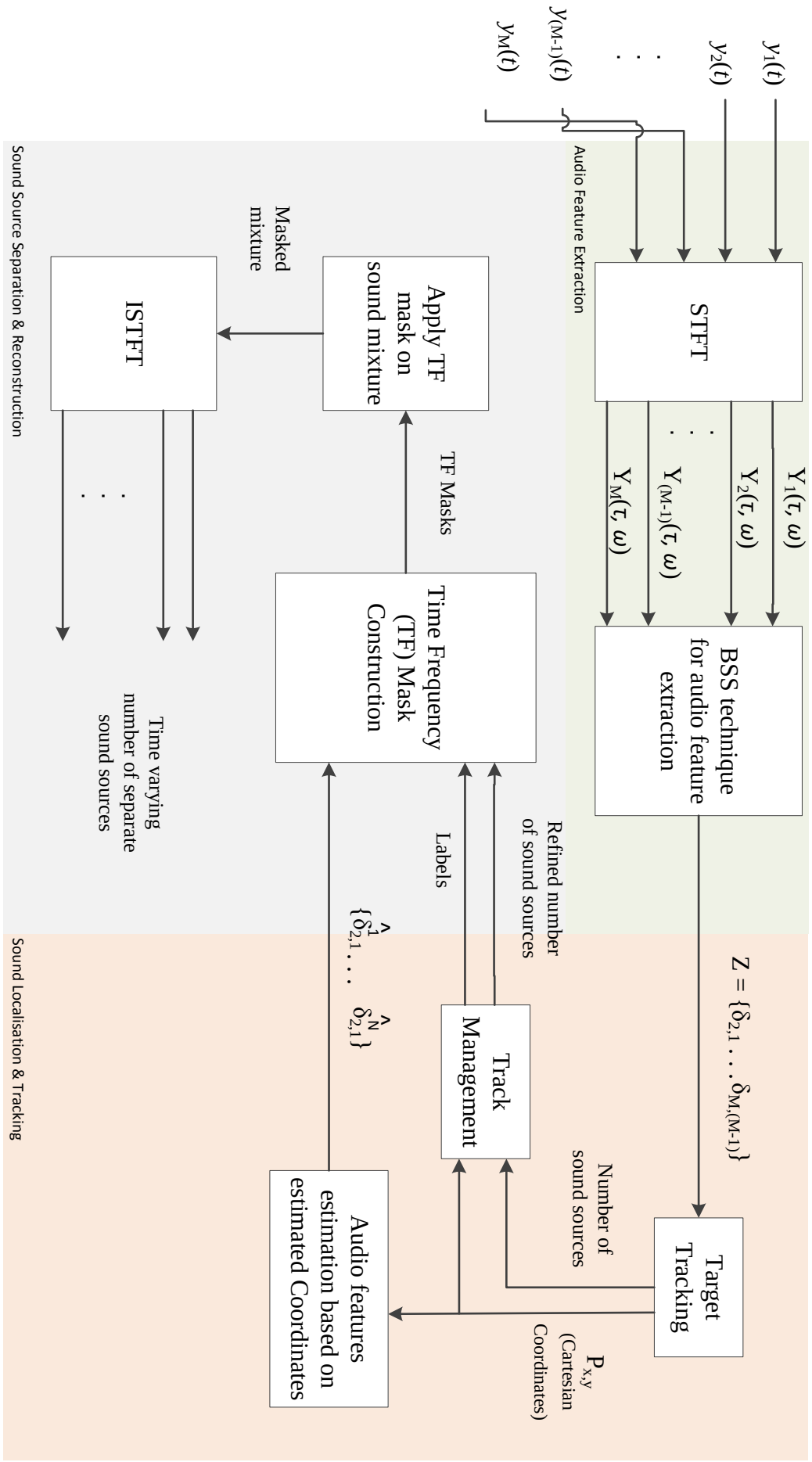


Figure 2.4: Block diagram of the Proposed Algrithm

2.7 Summary and Discussion

In summary, Chapter 2 has discussed the main problems related to the development of an online solution for localisation, tracking and separation of multiple speech sources. The sound source separation problem in the "conference room scenario" can essentially be broken down into the problems of sound source localisation, tracking and separation. For sound source localisation, the problem lies in extracting acoustic features from a sound source mixture. The problem of sound source tracking is the creation of tracks that link the locations and trajectories of the sound sources based on the observed acoustic features. The final problem of sound source separation lies in the data association of the tracks and estimating the mixing parameters from this information in order to separate the sound source mixtures. As a result of viewing the source separation problem in the "conference room scenario" from all three perspective, most techniques in the area of sound source separation, sound source localisation and sound source tracking were investigated.

The techniques that were investigated for sound source separation are HOS, SOS, BSS via signal sparsity and CASA. The source separation techniques using HOS and SOS are not utilized in this research as they are incapable of separating sound sources in underdetermined scenarios. CASA requires the modelling of the human hearing which would be too computationally complex for an online solution so this technique was ruled out. BSS via signal sparsity is the most viable technique as it is not computationally expensive to run and it is capable of extracting acoustic features in underdetermined scenario which suits the purpose of determining the acoustic features of time varying sound sources. These techniques are mainly used to extract the acoustic features of non-moving sound sources. With the changes in the acoustic features when the sound sources move, traditional sound source separation techniques are not enough to deal with the changes in the acoustic features. Sound tracking using blockwise batch processing proposes to process the signal of moving sound sources in blocks and to treat the moving sound sources are being pseudostationary within the time period of a block. By adapting this idea to be used in the sound source separation techniques, the problem of acoustic feature extraction for moving sound sources can be overcome.

The source localisation techniques reviewed are GCC, MUSIC, and SRP-PHAT.

MUSIC was ruled out as a source localisation technique as this technique was originally developed to handle narrowband signals so they would not perform as well in localising speech sources which are broadband signals. GCC-PHAT is a robust localisation technique even in reverberant scenario but it requires modifications in order to localisation multiple sound sources. SRP-PHAT overcomes the problem faced by GCC and a steered beamformer in terms of source localisation but it was not used in this research due to the high computational complexity required to run it. Furthermore, source separation techniques have been applied to achieve sound source localisation so it is preferable to use a sound source separation technique to extract the acoustic features required to localise the sound sources as the information generated by the techniques can later be used to achieve source separation.

Sound source tracking is a very challenging problem. Bayesian tracking is a common approach in dealing with target tracking problems. It creates a spatio-temporal relationship between the measured TDoA and the trajectory of the target tracked. An early work in using the Bayesian technique for sound source tracking is the use of PF to track moving source by Ward and Lehmann [64]. The limitation of this technique is the number of targets tracked so the work was further extended through the use of RFS techniques to track multiple targets in [77] and [88]. The PHD and CPHD filters used were not initially designed to handle nonlinear scenarios so these two filters require modifications in order to handle the dynamics model and measurement model of speech sources which are nonlinear. Moreover, the state extraction from particle clusters of the PHD filters require clustering so it is an expensive technique to run. Hence, these two RFS approaches were not used in this research which aims to provide and online solution for tracking and separating speech sources. The work presented in [66] has further extended the multiple sound source tracking to multiple sound source tracking and separation. The limitations of this research are the way in which it deals with the sound source presence as well as the data association between the different sound sources across time. Apart from that, the source sources were only tracked in one dimension. Multimodal sound source tracking and separation techniques were also reviewed in Chapter 2. This research requires an online recursive method to track and separate sound sources but multimodal techniques are either too computationally intensive or they require training data. Hence, the multimodal techniques reviewed were

not used as part of the proposed solution.

Based on the literature review of the techniques in source localisation, tracking and separation, it is believed that a source separation technique that can also be used for source localisation will be an ideal candidate to be used for acoustic feature extraction in this research. With the additional requirement of extracting acoustic features from time varying number of sound sources, a source separation technique which is capable of handling underdetermined scenarios is required. BSS via signal sparsity is a technique that meets both these requirements. With regards to multiple sound source tracking, the aim is to localise and track the time varying number of sound sources based on the extracted acoustic features. An RFS technique would be suitable as it is an elegant and mathematically tractable online solution that is capable of recursive multi-target tracking. Sound source separation of multiple moving sound sources can be achieved by utilising the track information to generate source labels and mixing parameters. The proposed solution is to exploit signal sparsity of received signals in the TF domain to extract the acoustic features and track the speech sources based on the extracted acoustic features using an RFS method. The details of the techniques used for acoustic features extraction and sound source tracking are discussed in Chapter 3 and Chapter 4 respectively. The main contribution of this thesis which is the integration of these two techniques to achieve sound source localisation, tracking and separation of moving speech sources will be presented in Chapter 5.

Chapter 3

Audio Feature Extraction via Signal Sparsity

Todays life comes before tomorrows.

– Haruto Souma

3.1 Overview

In a "conference room scenario", the speech sources are moving and the acoustic features change with the position of the speech sources. As the speech sources are defined by their unique acoustic features, the acoustic features of these speech sources across time needs to be known in order to perform source separation. In Chapter 3, the problem to be focused on is the extraction of acoustic features from a sound mixture signal containing a time varying number of moving speech sources. The aim of acoustic feature extraction in a "conference room scenario" is to extract the time varying attenuation, $a_{mn,x_{t,n}}$ and delay, $\delta_{mn,x_{t,n}}$, from the received speech source mixture. The acoustic features extraction process is illustrated in diagram 3.1.

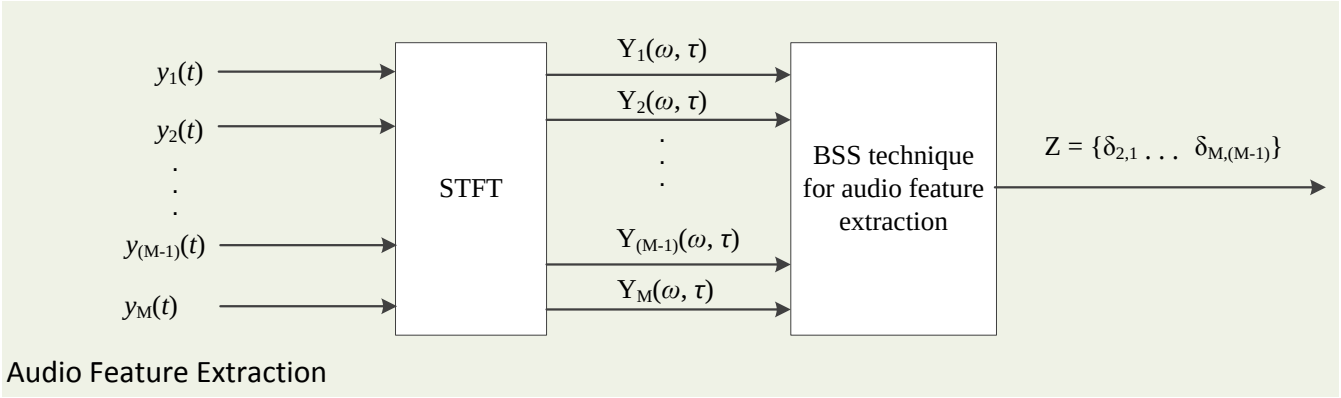


Figure 3.1: Block diagram of the Acoustic Feature Extraction Process

As mentioned earlier in Chapter 2, there are four main approaches to solving the BSS problem in an acoustic scenario - Higher Order Statistics (HOS), Second Order Statistics (SOS), Computer Auditory Scene Analysis (CASA) and BSS via signal sparsity. HOS, SOS and CASA are not applicable in the extraction of acoustic features for a time varying number of moving speech sources due to their limitations. HOS and SOS assume that the number of sound sources to be separated is less than the number of received signals while CASA requires a lot of computational power as it models the algorithm based on human hearing. Thus, the approach taken by this research is to exploit the sparsity of speech signals in the time frequency domain to achieve sound source localisation and separation.

The principle behind source separation via signal sparsity and how this principle can be applied to extract the acoustic features is discussed in this chapter. The suitability of applying BSS via signal sparsity in a "conference room scenario" is also analysed. The work in [97] has shown that the acoustic features extracted using a BSS via signal sparsity method can be used for source localisation purposes. This is due to the fact that BSS via signal sparsity is capable of extracting audio features in overdetermined, determined and underdetermined audio scenarios. Compared to other scenarios such as HOS, SOS and CASA which are only capable of handling overdetermined [98] and determined scenarios, BSS via signal sparsity is highly favourable when the number of sound sources to be tracked is unknown *a priori* and changes with time and the proposed algorithm has to be performed online.

Motivation

One of the earlier works that proposes the use of signal sparsity in the TF domain to separate sound sources is Degenerate Unmixing Estimation Technique (DUET) that was developed by Yilmaz and Rickard [4] in 2004. Since its introduction, it has spurred much research in the field of source separation via source sparsity exploitation. DESPRIT [99] [100], MENUET [101] and BSS combined with fuzzy C means clustering [102] are among the research that utilizes the theory introduced by DUET.

DESPRIT is an extension of DUET that combines DUET with ESPRIT. The motivation for such a combination was to perform blind signal separation on more than one signal mixture [100]. DESPRIT is capable of extracting acoustic features from an array of sensors whereas DUET only works for a single pair of sensors. DESPRIT also works under a more relaxed W-disjoint orthogonality assumption and is capable of handling echoic mixtures [100]. Despite the strengths of DESPRIT, it was not used as the acoustic feature extraction technique as it requires uniform linear array of sensors and it suffers from front back confusions [102]. DESPRIT also adds an additional layer of complexity to the acoustic feature extraction as the function played by the ESPRIT technique in this hybrid can be replaced by a target tracking technique to jointly estimate the number of sound sources and the location of the sound sources.

The sensor array limitation of DESPRIT is addressed by MENUET [101]. MENUET [101] is another form of DUET that utilizes multiple sensors instead of just a single pair. In MENUET, k-means clustering technique is used to cluster the acoustic features extracted. The number of clusters represent the number of sound sources. This allows an arbitrary number and array formation of microphones to be used as the location of the microphones need not be known beforehand. This algorithm is further extended in [102] and [103] through the use of a fuzzy C means clustering technique instead of a k-means clustering technique. The basic idea behind these two methods remain the same but the use of a soft clustering technique (fuzzy C means clustering) has an edge over a hard clustering technique (k-means clustering). The advantage of using a soft clustering technique is obvious in the sound quality of the reconstructed sound sources after separation [102]. The use of a hard clustering technique such as the k-means clustering produces a binary mask which will inevitably result in a lot of "musical tones" appearing in the separated sound sources. On the other hand, a soft clustering

technique like the fuzzy c means is able to mitigate the effects of "musical tones" and produces separated sound sources which sound more natural to human speech [104] .

Despite the merits of MENUET with the clustering techniques, MENUET is not used as an acoustic feature extraction technique in this research in which the number of sound sources are unknown and time-varying as the clustering techniques require prior knowledge of the sound source number and also the mean of these clusters. The mean of these clusters are usually estimated through an recursive procedure which cannot be guaranteed to converge [103]. Even if an unsupervised clustering technique such as [105] is used, it only adds computational complexity as the focus at this stage of research is to extract acoustic features from the sound mixture and the role of estimating the number of sound sources and refining the acoustic estimates can be better fulfilled by target tracking techniques. Apart from that, MENUET with the hard clustering technique is affected by outliers in the data and the centroids of the clusters will not necessarily be the true acoustic features of the individual sounds as these techniques might not converge [103]. Another limitation of using such clustering techniques is the limitation of the sound mixture to be comprised of non-moving sound sources. The clustering techniques are run recursively and each iteration shapes the clusters that represent the sound sources. As a result of this, MENUET with clustering techniques is unsuitable for online implementation as an online implementation of an acoustic feature extraction technique will require the acoustic features to be extracted in short segments of the source mixture. Due to these various limitations of clustering techniques, it is difficult to apply these clustering techniques to separate moving speech sources with an unknown number which changes with time. Hence, a new approach needs to be taken in order to separate the time-varying number of moving speech sources in the "conference room scenario".

In this research, it is proposed that the principal ideas introduced in DUET to be adapted and used for acoustic features extraction and source separation. As these ideas form a central part of the proposed solution, the background information on DUET is discussed. The main motivation of exploiting the signal sparsity in the TF domain stems from its flexibility as it can be integrated with other techniques to improve the separation performance. This is evident in [30] and [31] whereby the idea of signal sparsity is used and integrated with CASA techniques to improve the performance of

source separation using CASA. The use of fuzzy C means clustering to cluster the acoustic features in MENUET [102] further reinforces the idea that the core principle of exploiting signal sparsity to achieve source separation can be integrated with other techniques to enhance the performance of the algorithm. Furthermore, BSS via signal sparsity is capable of extracting acoustic features with low computational complexity in overdetermined, determined and underdetermined scenarios [106]. This allows the algorithm to extract the acoustic features in a "conference room scenario" as the number of speakers in the room are not known *a priori* and time-varying. Acoustic feature extraction can still be performed even if at any given instance the number of active sound sources exceeds the number of microphones available in the room. Apart from that, the information processed through the use of BSS via signal sparsity can be further exploited to separate the speech source mixture unlike other algorithms such as GCC or beamforming which are purely used for source localisation purposes.

3.2 Speech Source Separation via Signal Sparsity

DUET [4] is a robust and efficient BSS technique as long as the condition of *approximate W-disjoint orthogonality* is fulfilled. By exploiting the sparseness in the speech signals, DUET is able to separate sound source mixtures into their respective sources. There are three main stages to DUET - acoustic feature extraction, mask estimation and source separation.

3.2.1 Approximate W-disjoint orthogonality

The concept of *W-disjoint orthogonality* is explored in [107]. Two signals are considered as *W-disjoint orthogonality* if the windowed Fourier transform of those two signals are disjoint. Source signal sparsity can be referred to a situation whereby only one source is active at any time frequency point. It can be expressed as

$$S_n(\tau,\omega)S_o(\tau,\omega) = 0, \forall\tau,\delta,\forall n \neq o \tag{3.1}$$

with S_n and S_o being the Short Time Fourier Transform (STFT) of the respective observed speech signals, τ being the time and ω defined as the frequency.

The concept of *W-disjoint orthogonality* is restrictive in a real speech scenario as

the time frequency representation of the speech signals will still overlap when multiple speakers are speaking simultaneously. Hence, this condition is further relaxed to *approximate W-disjoint orthogonality* in [108] by exploiting the sparsity in the Gabor expansion of speech signals. The fact that large Gabor coefficients are usually concentrated in small percentage of the time frequency regions means that different signals will rarely have large Gabor coefficients in the same time frequency region. Hence, the approximate W-disjoint orthogonality is achievable by speech signals although the result of source separation still relies on the degree of approximate W-disjoint orthogonality between the speech signals. Clearer separation is made possible with a higher degree of approximate W-disjoint orthogonality.

With *approximate W-disjoint orthogonality*, sound separation for an arbitrary number of sound sources is possible as long as the sound sources do not overlap too much in the time-frequency domain. Such assumption holds true for most cases of speech as human speakers take turns to speak in most given scenarios [36]. However, the assumption of approximate W-disjoint orthogonality will be violated and the separation performance deteriorates when the room reverberation increases due to the effects of multipath.

3.2.2 Acoustic Feature Extraction

The first stage of DUET is acoustic feature extraction. As DUET exploits the approximate W-disjoint orthogonality of the speech signals in order to achieve source separation, STFT is performed on both of the observed sound mixtures. This is done in order to transform the observed mixtures into the time frequency domain in which the source signals are sparse. The signal model used in DUET is the anchoic signal model shown in equation (2.5)

$$y_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \delta_{mn}).$$

The STFT of the signals is defined as

$$F^W[s_n](\tau, \omega) \colon = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{-\infty} W(t-\tau) s_n(t) e^{-j\omega t} dt \tag{3.2}$$

with $F^W[s_n](\tau, \omega)$ being the STFTed source and W referring to the window function that is used for the STFT. For the sake of brevity, the rest of thesis will simply refer

to $F^W[s_n](\tau,\omega)$ as $S(\tau,\omega)$. The original DUET algorithm uses a Hamming window with the size of 1024 for the STFT as it gave the best performance among the different window functions that were compared [4]. After the sound mixtures have been transformed into the time frequency domain, the ratio of the mixtures is used to obtain the features - instantaneous attenuation and instantaneous delay. The time frequency ratio, $\mathcal{H}_{21}$, can be expressed as

$$\mathcal{H}_{21} = \frac{Y_2(\tau,\omega)}{Y_1(\tau,\omega)} = \frac{\sum_{n=1}^{N} a_n e^{-j\omega\delta_n} S_n(\tau,\omega)}{\sum_{n=1}^{N} S_n(\tau,\omega)}. \tag{3.3}$$

By exploiting the approximate W-disjoint orthogonality which states that only one sound source will have significant contribution in a particular time frequency point, equation (3.3) can be further simplified to

$$\mathcal{H}_{21} = \frac{a_n e^{-j\omega\delta_n} S_n(\tau,\omega)}{S_n(\tau,\omega)} = a_n e^{-j\omega\delta_n}. \tag{3.4}$$

In the original DUET system, both the instantaneous attenuation and the instantaneous delay are calculated for each time frequency point of the sound mixture and then clustered by a power weighted histogram. The instantaneous attenuation and the instantaneous delay can be obtained from the time frequency ratio, $\mathcal{H}_{21}$, based on equations (3.5) and (3.6)

$$a_n = |\mathcal{H}_{21}| \tag{3.5}$$

$$\delta_n = \frac{1}{-\omega} \angle \mathcal{H}_{21} \tag{3.6}$$

with $|\mathcal{H}_{21}|$ referring to the absolute value of the time frequency ratio while $\angle\mathcal{H}_{21}$ refers to the angle of the time frequency ratio. In the original algorithm, the instantaneous attenuation, a_n, is not used in the construction of the power weighted histogram. Instead, the symmetric attenuation, α_n is used. The symmetric attenuation, α_n, can be calculated from the instantaneous attenuation, a_n, based on the following formula:

$$\alpha_n := a_n - \frac{1}{a_n} \tag{3.7}$$

The symmetric attenuation, α_n, is used in place of the instantaneous attenuation, a_n, in order to have sound source's attenuation reflect symmetrically at the centre point ($\alpha_n = 0$) even if the microphone's signals are swapped. If the instantaneous attenuation was used, a, a swap of the microphone signals will result in the change from a_n to $\frac{1}{a_n}$. With

the use of the symmetric attenuation, a swap of the microphone signals will only result in a change from α_n to $-\alpha_n$ without a change in the value of α_n [36]. Sound sources which are louder on microphone 1 will have $\alpha_n < 0$ while sound sources which are louder on microphone 2 will have $\alpha_n > 0$, with the use of the symmetric attenuation, α_n [36].

A power weighted histogram is constructed to cluster the estimated acoustic features into groups which define the original sound sources. The power weighted histogram is influenced by Maximum Likelihood algorithm [4]. The power weighted histogram is a valid method of clustering the acoustic features based on three observations by the authors of DUET [4]:

- *Observation 1: Most of the source energy is captured within a small rectangle of instantaneous acoustic features centered on the true acoustic features*
- *Observation 2: Observation 1 holds true even for individual sound sources in the sound source mixture*
- *Observation 3: The smoothed 2D-weighted histogram of the instantaneous acoustic features is in one-to-one correspondence with the rectangle centers in Observation 2*

The authors of DUET chose to use the weighted estimates as they are more accurate estimates of the true acoustic features [36]. The power weighted attenuation can be expressed as

$$\tilde{\alpha_n} := \frac{\int\int_{(\tau,\omega)\in\Omega_n} |Y_1(\tau,\omega)Y_2(\tau,\omega)|^p \omega^q \tilde{\alpha}(\tau,\omega)\mathrm{d}\tau\mathrm{d}\omega}{\int\int_{(\tau,\omega)\in\Omega_n} |Y_1(\tau,\omega)Y_2(\tau,\omega)|^p \omega^q \mathrm{d}\tau\mathrm{d}\omega} \tag{3.8}$$

while the power weighted delta estimate can be expressed as

$$\tilde{\delta_n} := \frac{\int\int_{(\tau,\omega)\in\Omega_n} |Y_1(\tau,\omega)Y_2(\tau,\omega)|^p \omega^q \tilde{\delta}(\tau,\omega)\mathrm{d}\tau\mathrm{d}\omega}{\int\int_{(\tau,\omega)\in\Omega_n} |Y_1(\tau,\omega)Y_2(\tau,\omega)|^p \omega^q \mathrm{d}\tau\mathrm{d}\omega} \tag{3.9}$$

The time frequency weight of a given time frequency point is $|Y_1(\tau,\omega)Y_2(\tau,\omega)|^p\omega^q$ whereas Ω_n refers to a set of (τ,ω) points (determined to be associated with the nth cluster). Various choices for p and q hold for different circumstances:

- $p = 0$, $q = 0$ was the counting histogram used in original DUET [107]. The counting histogram does not consider the speech energy nor the speech frequency.

- $p = 1$, $q = 0$ was motivated by the Maximum likelihood symmetric attenuation [4].
- $p = 1$, $q = 2$ was motivated by the Maximum likelihood delay estimator [4].
- $p = 2$, $q = 0$ was proposed to reduce delay estimate bias [4].
- $p = 2$, $q = 2$ is a good choice for low SNR scenarios or speech mixtures [36]

$p = 1$ and $q = 0$ is a good default choice for most applications but $p = 2$ and $q = 2$ is best suited for speech signals [36]. The weighting of $p = 2$ and $q = 2$ helps to accentuate the contributions by speech which have a higher energy and frequency content compared to background noise. The choice of $p = 2$ and $q = 2$ has the limitation of failing to pick up the a sound source if the sound source's energy contribution is significantly lower than the others. For the purpose of sound separation in a full sound signal, this will be less of a problem as the the contributions of all the sound sources in the full signal will be more or less equal but this will not be the case when the TF-weighted histogram is examined on a frame by frame basis. The pairing of $p = 0.5$ and $q = 0$ is suggested when the sound sources are not of equal power as it prevents the dominant sound sources from obscuring the peaks of the smaller sound sources in the power weighted histogram [36].

The power weighted histogram can be used to estimate the mixing parameters as the instantaneous parameters will cluster around the true parameters. Hence, the number of sound sources can be represented by the number of peaks in the power weighted histogram while the mixing parameters - symmetric attenuation and relative delay, are represented by the center of these peaks. An example of a power weighted histogram featuring a 30dB source mixture with sources originating from 45 degrees left and 60 degrees right of the microphone pair is shown in figure 3.2. The two peaks in the power weighted histogram indicates that there are two sound sources. The highlighted peaks shows the mixing parameters, ($\tilde{\alpha} = 0.041$ and $\tilde{\delta} = -1.556$) for the sound source emanating from 60 degrees right of the microphone pair and ($\tilde{\alpha} = -0.041$ and $\tilde{\delta} = 1.273$) for the one emanating from 45 degrees left of the microphone pair. The estimated acoustic features from the peaks are used for the time frequency masks' construction.

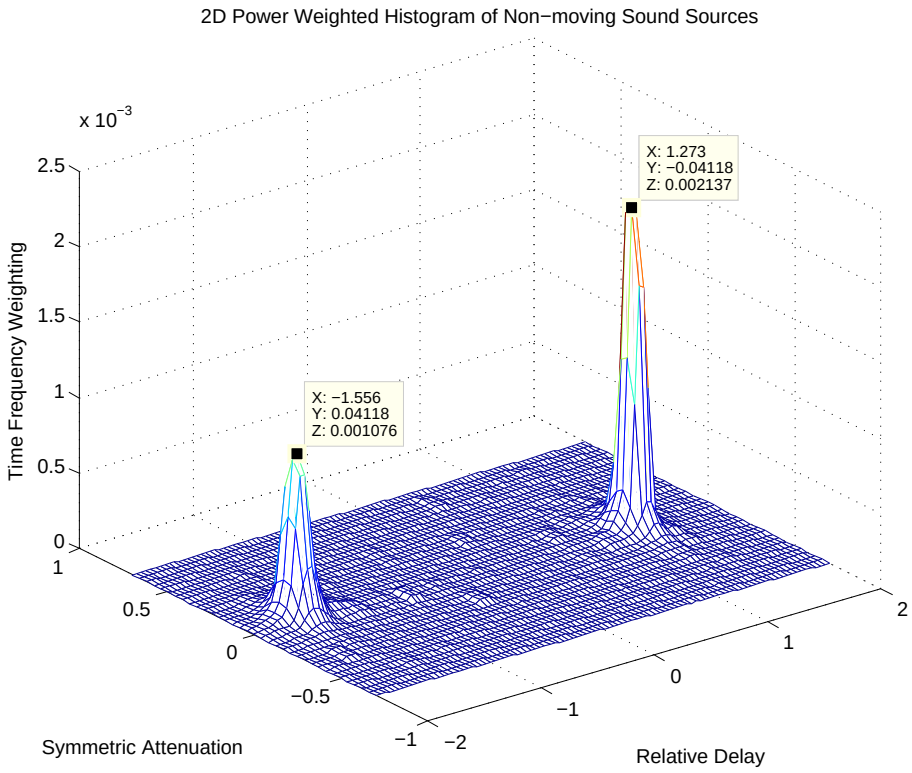


Figure 3.2: An example of a power weighted histogram for a mixture of two non-moving sound sources

3.2.3 Mask Construction and Source Separation

The mask used for source separation is constructed using the acoustic features estimated from the power weighted histogram. The estimated attenuation, $\tilde{a_n}$, is used in the mask construction so the symmetric attenuation, $\tilde{\alpha_n}$, is converted back to the attenuation using equation 3.10

$$\tilde{a_n} = \frac{\tilde{\alpha_n} + \sqrt{\tilde{\alpha_n^2}}}{2} \tag{3.10}$$

A scoring system is devised by using the instantaneous likelihood function to measure the closeness of a particular time frequency point to the peak value [4]. A peak is assigned to each time-frequency point using equation ([36]):

$$J(\tau,\omega) := \arg\min_{o} \frac{|\tilde{a_o} e^{-j\omega\tilde{\delta_o}} Y_1(\tau,\omega) - Y_2(\tau,\omega)|^2}{1 + \tilde{a_o}^2} \tag{3.11}$$

A value of 1 is assigned to acoustic feature estimates which belong to a particular peak while 0 is assigned to the other parameters. The binary mask can be expressed as:

$$\tilde{M_n}(\tau,\omega) := \begin{cases} 1 & J(\tau,\omega) = n \\ 0 & otherwise \end{cases} \tag{3.12}$$

With the assumption made that the audio features are mutually exclusive and only belong to a particular sound source, the masking technique used is a hard masking technique. The binary mask of each sound source is applied to the sound source mixture using equation ([36])

$$\tilde{s_n}(\tau,\omega) = \tilde{M}_n(\tau,\omega)\left(\frac{Y_1(\tau,\omega) + \tilde{a_n}e^{-j\omega\tilde{\delta_n}}Y_2(\tau,\omega)}{1+\tilde{a_n}^2}\right) \tag{3.13}$$

The final step in DUET involves performing Inverse STFT on the estimated sound signal to reconstruct the original sound sources in the time domain. The DUET technique is summarized in the following algorithm[36]:

Algorithm 2 Pseudocode for DUET

Acoustic Feature Extraction

1: Transform the received signals, $y_1(t)$ and $y_2(t)$, into their their time-frequency representation, $Y_1(\tau,\omega)$ and $Y_2(\tau,\omega)$ via STFT

2: Calculate the time frequency ratio, $\mathcal{H}_{21} = \frac{Y_2(\tau,\omega)}{Y_1(\tau,\omega)}$ and extract the instantaneous attenuation, $a = |\mathcal{H}_{21}|$, and instantaneous delay, $\delta = \frac{1}{-\omega}\angle\mathcal{H}_{21}$, from it.

Acoustic Feature Clustering

3: Construct the TF-weighted histogram

4: Locate the peaks in the histogram and determine the value of the peak centers as the peak centers determine the mixing parameters, $(\tilde{\alpha},\tilde{\delta})$

Mask Construction and Source Separation

5: Construct the TF binary masks based for each peak with their respective peak centers, $(\tilde{\alpha},\tilde{\delta})$

6: Apply each mask to the appropriately aligned mixtures

7: Convert the separated sound sources from the TF domain back to the time domain through ISTFT

3.3 Limitations of BSS via Signal Sparsity in Separation of Multiple Moving Speech Sources

BSS via signal sparsity has been shown to be a valid separation method when applied to scenarios where the number of sound sources are known and the sound sources are non-moving. BSS via signal sparsity relies on approximate W-disjoint orthogonality assumption to extract acoustic signals from a sound source mixture and clustering of these acoustic feature to achieve source separation. The main advantage of BSS via signal sparsity is its capability to extract acoustic features, a_{mn} and δ_{mn}, even in underdetermined scenarios due to the approximate W-disjoint orthogonality assumption

[4]. Furthermore, the computational complexity of a basic BSS via signal sparsity technique such as DUET is low, making it suitable for online applications. Techniques such as MENUET with fuzzy C clustering shows that the basic idea of exploiting signal sparsity for sound source separation can be integrated with other techniques to further improve the source separation performance. Despite these benefits offered by BSS via signal sparsity, there are limitations to these techniques when they are to be applied in a "conference room scenario".

The first limitation of applying BSS via signal sparsity in "conference room scenario" is the lack of information on the number of active sound sources. Apart from the power weighted histogram used in DUET, clustering techniques such as k-means clustering and fuzzy C means clustering requires the number of sound sources to be known in order to sort the acoustic features into their distinctive clusters. This information is unavailable in a "conference room scenario" as the number of sound sources are not known *a priori* and this number changes when the active sound sources appear or disappear with time. Acoustic features clustering with power weighted histogram does not require prior information on the number of active sound sources and the number of peaks which form indicate the number of sound sources. Hence, acoustic feature clustering using a power weighted histogram is a viable clustering method in a "conference room scenario".

The second limitation faced by BSS via signal sparsity when it is applied in "conference room scenario" is the amount of audio information required to separate the signals. Conventional BSS via signal sparsity techniques perform an analysis of the source signals across the whole measurement period. Due to the assumption that the acoustic features, a_{mn} and δ_{mn}, are constant, the longer analysis period will yield a better separation performance as the extra information helps to distinguish the true acoustic features of each sound source. Furthermore, the longer measurement period helps the In a "conference room scenario", the acoustic features, $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$ are dependent on the position of the sound source and are no longer constant throughout the whole measurement period. A moving sound source which has changing acoustic features will not be well represented by the clustering techniques used by conventional BSS via signal sparsity techniques. An example of this is shown when a power weighted histogram constructed upon a sound mixture of 2 moving speakers

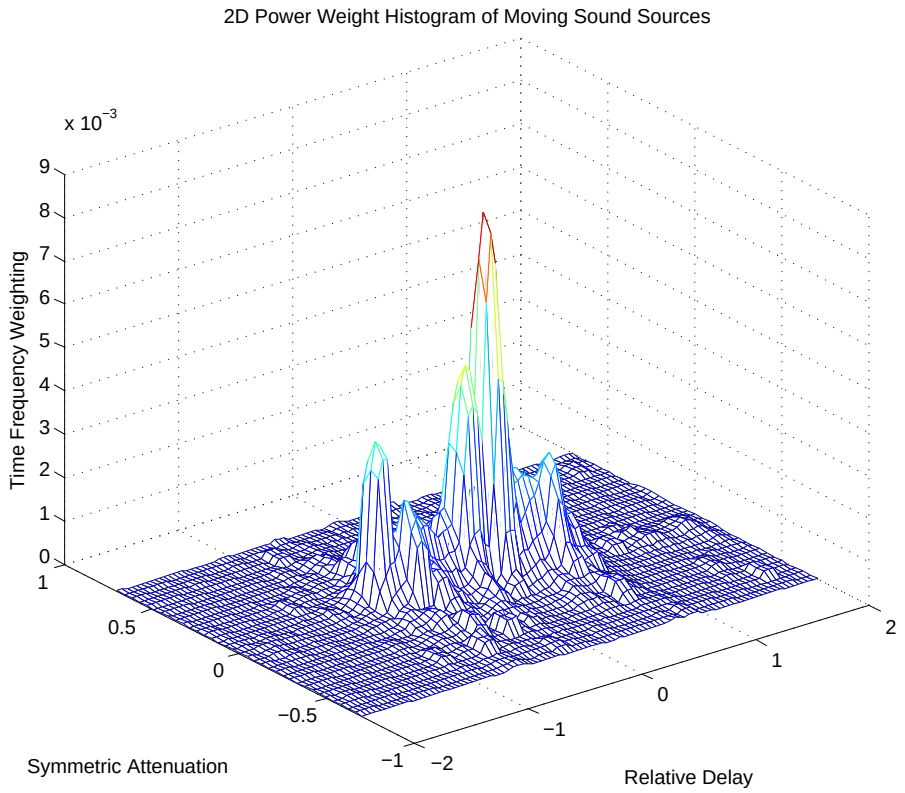


Figure 3.3: An example of a power weighted histogram for a mixture of two moving sound sources

with 30dB noise and 150ms reverberation time in figure 3.3. Clustering techniques used in conventional BSS via signal sparsity techniques such as the power weighted histogram and fuzzy C means clustering are unable to determine the means or peaks of the clusters which are used to estimate the true acoustic features if the acoustic features analysed are constantly changing. When the whole measurement period is analysed, the changing nature of the moving sound sources' acoustic features will cause these clusters to merge. This is shown in figure 3.3 where there are no clear peaks to indicate the unique clusters for each sound source in the mixture. Conventional BSS via signal sparsity techniques which rely on these clusters to separate the individual sound sources can no longer perform separation as there no unique clusters that identify the different sound sources.

A possible solution to overcome the aforementioned limitation is to analyse the sound source mixture on a frame-by-frame basis by exploiting the fact that the moving sound sources will remain pseudo-stationary within the time frame which they are inspected. By using the assumption that the acoustic features are pseudo-stationary within a short time frame, $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$ can be approximated to just a_{mn} and

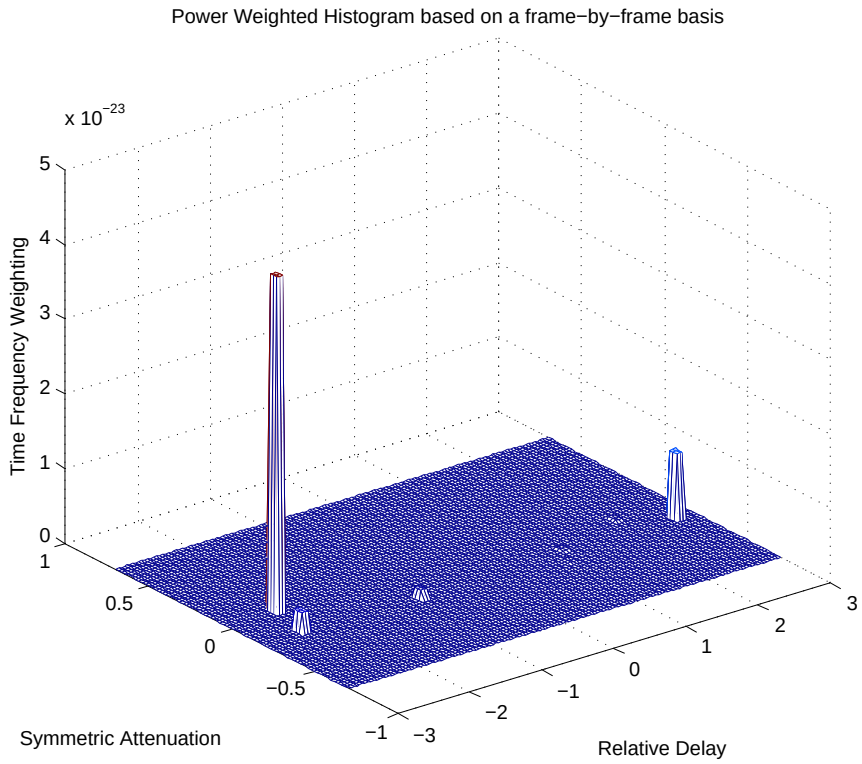


Figure 3.4: An example of a power weighted histogram for a mixture of two moving sound sources based on a frame by frame analysis

δ_{mn} within the time frame. This approximation allows clusters to be formed for each time frame and sound source separation to be performed. As the clusters are formed for each time frame, the number of sound sources are also determined for each time frame, making it possible to determine the number of active sound sources which is constantly changing across time.

However, this method of acoustic feature extraction for moving sound sources has its own shortcomings as well. Most clustering techniques require a few iterations to converge to the true mean or centroid so the short time period used for analysis does not contain enough data for these clusters to converge. As the cluster is constructed for a frame which is just a short period in time, the energy contribution of the sound sources are lower as well. Thus, contributions by noise will be relatively larger in the clusters causing spurious peaks to appear in the estimation [64]. An example of a power weighted histogram used to analyse the same sound mixture of 2 moving speakers with 30dB noise and 150ms reverberation on a frame by frame basis is shown in figure 3.4.

The application of a frame by frame analysis allows conventional BSS via signal

sparsity techniques to extract acoustic features from moving sound sources but it also leads to a new problem of identity ambiguity in the separated sound sources. Due to the greater contribution of noise and reverberation, other clusters or peaks resulting from the noise or reverberation will emerge. As BSS via signal sparsity rely on the clusters to identify the different sound sources, the additional clusters resulting from noise and reverberation will prevent the true acoustic features to be matched to their respective sources. As shown in figure 3.4, there are four peaks despite the fact that there are only two sound sources. This is an example of noise corrupting the acoustic feature estimates when the acoustic feature extraction process is performed on a frame by frame basis and the energy contribution of the sound source is low. These extra clusters due to noise and reverberation are clutters which cause identity ambiguity within the frame. The identity ambiguity problem further extends from frame to frame. Conventional BSS via signal sparsity techniques are incapable of associating the different acoustic features $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$ which have been approximated to just a_{mn} and δ_{mn}, from frame to frame. As the true acoustic features are constantly changing when the sound sources move, the mean or peak of the clusters will shift from frame to frame. Hence, a multi-target tracking technique with data association is required to solve this problem. Further details on multi-target tracking using RFS and its application in the "conference room scenario" will be discussed in Chapter 4.

3.4 Summary

In summary, BSS via signal sparsity can be used to extract the acoustic features $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$ from a sound mixture which contain moving targets. The justification of using BSS via signal sparsity stems from the simplicity of this method and its capability to separate more sound sources than the number of sensors available. The fundamentals of using signal sparsity to achieve sound source separation is also discussed in this chapter. Chapter 3 concludes with the discussion on the limitations of BSS via signal sparsity when it is applied to a "conference room scenario".

In order to tackle the problem of extracting acoustic features from a time varying number of sources, BSS via signal sparsity is a highly suitable method as it has been shown to be capable of separating more sound sources than the number of sensors

available [106]. As a result of this, acoustic feature extraction will not be a problem when the number of sound sources in a room exceeds the predetermined number of sensors in the room. The use of *W-disjoint orthogonality* assumption in acoustic feature extraction was analysed. The benefit of using this assumption is the low computational complexity it offers. By using the assumption of *W-disjoint orthogonality*, the computational complexity of the problem does not scale when there are multiple sound sources [4] but the accuracy of the acoustic feature estimates do suffer when there are no strong direct sound paths. Noise and reverberation will contribute additional sound paths which weaken the *W-disjoint orthogonality* assumption and affect the accuracy of the acoustic feature estimates.

BSS via signal sparsity suffers from two main limitations when it is applied in the "conference room scenario". The first limitation is the requirement of knowledge on the number of sound sources in order to cluster the acoustic features. As the number of active sound sources are not known a priori and this number is changing throughout the measurement period, clustering techniques such as k-means and fuzzy C means clustering are incapable of sorting the extracted acoustic features into clusters. A possible solution is to apply a power weighted histogram which does not require the number of sound sources to be known a priori. The second limitation is the amount of information and time required for the clusters to converge in conventional BSS via signal sparsity techniques. Clustering techniques such as k-means and fuzzy C means clustering need to run recursively in order for the clusters to converge around the true acoustic features. The acoustic features in the "conference room scenario" vary with time so the means of the clusters will be constantly shifting and clusters will merge. Without unique clusters to represent the individual sound sources in the source mixture, sound source separation cannot be achieved. This limitation can be partially achieved by analysing the source mixture on a frame by frame basis. By doing so, the acoustic features $a_{mn,x_{t,n}}$ and $\delta_{mn,x_{t,n}}$ can be approximated to just a_{mn} and δ_{mn} for each time frame. The problem that arises from this is identity ambiguity which arises from spurious peaks of noise and reverberation due to the relatively lower signal energy contribution in each time frame. The data ambiguity problem extends from frame to frame as conventional BSS via signal sparsity techniques are unable to associate the identity of the sound sources from frame to frame. A method to overcome this limitation is to use an RFS

multi-target tracking technique to filter out the spurious peaks and offer data association between the acoustic features in order to jointly estimate the time varying number of sound sources and track the sound sources.

Chapter 4

Speech Source Localisation and Tracking

Chase two birds and you will catch two birds.

– Tendou Souji

4.1 Overview

In Chapter 3, the method in which BSS via signal sparsity can be used to extract the acoustic features from a mixture of moving speech sources is discussed. In a "conference room scenario", the number of active sound sources and the acoustic features are time varying and conventional BSS via signal sparsity techniques are faced with several limitations when they are applied in the "conference room scenario". These limitations of conventional BSS via signal sparsity techniques can be overcome by performing a frame by frame analysis of the signal instead of analysing the whole signal measurement period. By doing so, the time varying acoustic features can be assumed to be constant for each time frame. The source separation problem in the "conference room scenario" is transformed into a multi-target tracking problem by performing BSS on frame by frame basis and extracting acoustic features for each of the time frames.

The problem of speech source separation in a "conference room scenario" is similar to the classical multi-target tracking problem. The number of active speech sources at any given time is random and these speech sources are constantly moving. By applying the target tracking technique in this context, the speech sources which can

appear or disappear across time can be viewed as the targets to be tracked and the acoustic features can be viewed as the measurements generated. The challenge of this problem lies in the need to not only track the evolution of the speech sources' individual target states in time but the change in the number of targets as well due to the births and deaths of targets [109]. The reliability of the microphones also come into question as the microphones may or may not detect the speech sources. Furthermore, even if the targets are detected, the reliability of the measurements received are also of concern as not all the measurements are generated by the active speech sources. Apart from the measurements generated by the active speech sources, the microphones may also receive clutter which is a set of spurious measurements. In speech source tracking, the clutter which corrupts the acoustic features extracted are the spurious peaks due to noise and reverberation. In target tracking problem, the other challenge is data association. The identity of the targets tracked are maintained throughout the whole tracking process. In speech source tracking, the data association problem arises when the extracted acoustic features have to be matched with their respective speech sources across the time frames. The speech reconstruction after the separation process relies on applying a TF mask made from the different acoustic features which belong to the same speech source.

Despite the similarities with classical target tracking applications, speech source tracking presents its own distinctive challenges. In a conversation, speakers tend to pause in between words and when they are taking turns to talk. The first challenge is the low rate of detection due to the silence period in speech signals. The low rate of detection makes the speech sources more difficult to track as the measurements are not constantly generated. Besides that, the low rate of detection makes it harder to distinguish the true speech sources from multipath generated by the speech sources. These silence periods present a different challenge during data association as the acoustic features after the silence period needs to be matched to the ones before the silence period. The difficulty lies in making the decision of assigning a new identity for the acoustic features generated after the silence period or to categorise it with a previous speech source that has stopped generating measurements. In multi-target tracking, the algorithm needs to jointly estimate the number of sound sources and the state of these

sound sources based on a set of measurements amid the presence of probability detection uncertainty, data association uncertainty and clutter [13]. When applied to track speech sources, this onerous problem is further compounded with the unique set of challenges presented by speech sources.

By casting the speech source separation problem in a multi-target tracking context, established target tracking method can be applied to solve it. Thus, the focus of Chapter 4 will be on the problem of jointly estimating the number of speech sources and tracking these speech sources based on a set of measurements comprising of the acoustic features extracted using an RFS multi-target tracking method. This process is illustrated in the block diagram 4.1.

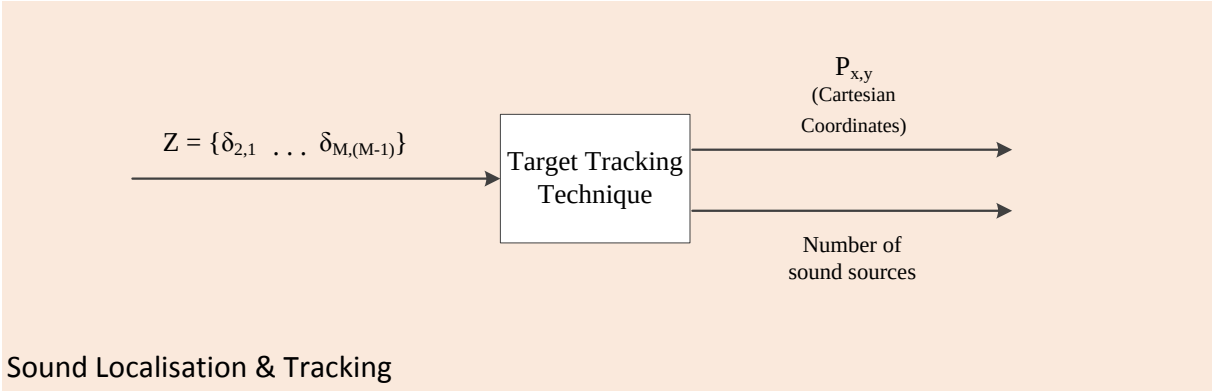


Figure 4.1: Block diagram of the Target Tracking Process

An overview of the various sound source tracking techniques using Random Finite Sets (RFS) has been given in subsection 2.5.2. The RFS methods used in [77] and [88] utilise RFS filtering methods which propagate the intensity of function - Probability Hypothesis Density (PHD) and Cardinalized Probability Hypothesis Density (CPHD) filters respectively. Since then, various developments have taken place in the field of RFS research and the Cardinality Balanced Multi-Target Multi Bernoulli (CBMeMBer) is one of the more advanced version of the RFS method which approximates the multi-target Bayes posterior density as a multi-Bernoulli RFS [13]. The Sequential Monte Carlo (SMC) implementation of the CBMeMBer filter is designed to track targets with nonlinear dynamics. Speech source tracking in a conference room scenario involves the tracking of speakers who are assumed to be moving in a nonlinear manner. Thus, the SMC implementation of the CBMeMBer filter is more suitable to be used in the tracking of speech sources compared to the PHD and the CPHD filter which were

initially designed to track targets with linear movements. In terms of computational complexity, the SMC implementation of the CBMeMBer filter is the same as the PHD filter which scales linearly with the increase in the number of targets but it is less complex than the CPHD which scales linearly in the number of targets but cubic in the number of measurements [13]. Under certain scenarios, the SMC-CBMeMBer filter has been shown to outperform the capabilities of the SMC-PHD and the SMC-CPHD filter despite a having a lower computational complexity [13]. As a result of this, the multi-target tracking technique used in this research is the CBMeMBer filter.

4.2 Cardinality Balanced Multi-Target Multi Bernoulli (CBMeMBer) Filter

The Cardinality Balanced Multi-target Multi-Bernoulli (CBMeMBer) [13] filter which was developed to overcome the cardinality bias of the MeMBer filter [110] is the multi-target tracking technique used in this research. In CBMeMBer filter, the multi-target posterior density is approximated as a multi-Bernoulli RFS. Most RFS filters (including CBMeMBer) operate under the following modelling assumptions [73, p.76]. The first assumption made is the independence of the measurements generated by each target. Secondly, the targets births are assumed to follow a multi-Bernoulli RFS independent of target survivals. Finally, the clutter is assumed to be distributed by a Poisson RFS and it is independent of the measurements generated by the tracked targets. In the case of CBMeMBer, it is further assumed that the measurements received have a lower rate of false positives and false negatives.

4.2.1 Random Finite Set Models

An overview of the RFS models used by CBMeMBer is required in order to have a better understanding of the CBMeMBer. An RFS is a random collection of states in the finite space which is characterised by its probability distribution [90]. The two main RFS models used by the CBMeMBer is the Poisson RFS and the multi-Bernoulli RFS. The Poisson RFS is used to model the clutter whereas the multi-Bernoulli RFS is used as an approximation to the full multi-target Bayes posterior density.

4.2.1.1 Poisson RFS

A Poisson RFS on $\mathcal{X} \subseteq \mathbb{R}^n$ is defined by its intensity function, $\mathfrak{v}$, which is also known as the Probability Hypothesis Density (PHD). The defining characteristic of a Poisson RFS is its cardinality which is Poisson distributed [13]. The mean, N, of the Poisson RFS is an integral of the intensity function:

$$\bar{N} = \int \mathfrak{v}(x)dx \tag{4.1}$$

For a given cardinality, the elements, x of the Poisson RFS X are each independently and identically distributed according to $\frac{\mathfrak{v}(\cdot)}{\bar{N}}$ [12]. The probability density of the Poisson RFS can be expressed as:

$$\pi(X) = e^{\bar{N}}\mathfrak{v}^X \tag{4.2}$$

The probability density of the Poisson RFS is used to model the clutter.

4.2.1.2 Multi-Bernoulli RFS

In the CBMeMBer filter, a multi-Bernoulli RFS is used to approximate the multi-target Bayes posterior density. For a single Bernoulli RFS, it has the probability, r, to be a unit set with only an element distributed according to the probability density, p or it has the probability, $1 - r$ of being empty. The probability density of a Bernoulli RFS can be defined as

$$\pi(X) = \begin{cases} r.p(x) & X = \{x\} \\ 1 - r & X = \emptyset. \end{cases} \tag{4.3}$$

The cardinality of a Bernoulli RFS is a Bernoulli distribution with the parameter r. The multi-Bernoulli RFS is just the union of a finite number of independent Bernoulli RFS

$$X = \cup_{i=1}^{\mathbb{M}} X^{(i)} \tag{4.4}$$

with $\mathbb{M}$ being the number of independent Bernoulli RFS, $X^{(i)}$ which has the existence probability, $r^{(i)}$ and the probability density, $p^{(i)}$. The probability density of a multi-Bernoulli RFS can be described by [110]

$$\pi(X) := \begin{cases} \Pi_{j=1}^{\mathbb{M}}(1 - r^{(j)}) & X = \emptyset \\ \Pi_{j=1}^{\mathbb{M}}(1 - r^{(j)}) \sum_{1 \leq i_1 \neq, \cdots, \neq i_N \leq \mathbb{M}} \Pi_{j=1}^{n} \frac{r^{(i_j)} p^{(i_j)(x_j)}}{1 - r^{(i_j)}} & X = \{x_1, \cdots, x_N\} \end{cases} \tag{4.5}$$

The rest of this thesis will use the abbreviated form of the full multi-Bernoulli density in equation (4.5), expressed as $\pi = \{r^{(i)}, p^{(i)}\}_{i=1}^{\mathbb{M}}$.

4.2.1.3 Multi-target System in CBMeMBer

In Section 2.5.2, it was shown that multi-target tracking systems operate in the multi-target space, $\mathcal{F}(\mathcal{X})$, and the multi-target observation space, $\mathcal{F}(\mathcal{Z})$. For a given time k, the state within the multi-target set, $x_{k-1} \in X_{k-1}$ can cease to exist with the probability $1 - p_{S,k}(x_{k-1})$ or survive with the probability $p_{S,k}(x_{k-1})$ and move to a new state with the transition density $f_{k|k-1}(x_k|x_{k-1})$. As mentioned earlier in equation (2.31), the multi-target state, X_k, at a given time k can be defined as

$$X_k = \left[\bigcup_{x_{k-1} \in X_{k-1}} S_{k|k-1}(x_{k-1}) \right] \cup \Gamma_k \tag{4.6}$$

In CBMeMBer, the dynamics of each target is described with a Bernoulli RFS. Hence, $S_{k|k-1}(x_{k-1}$ is a multi-Bernoulli RFS defined by the parameters $r = p_{S,k}(x_{k-1})$ and $p(x_k) = f_{k|k-1}(x_k|x_{k-1})$ while Γ_k is a multi-Bernoulli RFS that models the spontaneous births of the targets [13].

At a given time, k, a target, $x_k \in X_k$ can be detected with the probability $p_{D,k}(x_k)$ or missed by the probability $1-p_{D,k}(x_k)$. A detected target has the likelihood, $g_k(z|x_k)$, of generating an observation, z_k. Besides the true measurements generated by the target, clutter or false alarms, K_k, with the intensity function, $\kappa_k(\cdot)$ might be detected by the sensors. As shown in equation (2.31), the multi-target measurement, Z_k, can be expressed as

$$Z_k = \left[\bigcup_{x \in X_k} \Theta_k(x) \right] \cup K_k \tag{4.7}$$

In CBMeMBer, $\Theta_k(x)$ is a Bernoulli RFS which describes the measurement generated by the target at time k. $\Theta_k(x)$ is defined by the parameters $r = p_{D,k}(x_k)$ and $p(\cdot) = g_k(\cdot|x_k)$. The clutter, K_k, is modelled as a Poisson RFS in CBMeMBer.

4.2.2 CBMeMBer Recursion

Due to the computational complexity involved in the propagation of a full multi-target Bayes shown in equation (2.33) and (2.34), the CBMeMBer propagates a MeMBer approximate which comprises of a finite and time-varying number of hypothesized

tracks. These tracks are defined by their probability of existence and the probability density of each respective track [13]. The CBMeMBer filter has two main stages: **prediction** and **update**.

During the prediction stage, the prior multi-target density at time, k, is predicted based on the posterior multi-target density at time, $k-1$. If that posterior multi-target density is a multi-Bernoulli in the form of [13]:

$$\pi_{k-1} = \{(r_{k-1}^{(i)}, p_{k-1}^{(i)})\}_{i=1}^{\mathbb{M}_{k-1}} \tag{4.8}$$

the predicted multi-target density will also be a multi-Bernoulli in the form of

$$\pi_{k|k-1} = \{(r_{P,k|k-1}^{(i)}, p_{P,k|k-1}^{(i)})\}_{i=1}^{\mathbb{M}_{k-1}} \cup \{(r_{\Gamma,k}^{(i)}, p_{\Gamma,k}^{(i)})\}_{i=1}^{\mathbb{M}_{\Gamma,k}} \tag{4.9}$$

whereby the existence probability and the probability density of the state are:

$$r_{P,k|k-1}^{(i)} = r_{k-1}^{(i)} \langle p_{k-1}^{(i)}, p_{S,k} \rangle, \tag{4.10}$$

$$p_{P,k|k-1}^{(i)}(x) = \frac{\langle f_{k|k-1}(x|\cdot), p_{k-1}^{(i)} p_{S,k} \rangle}{\langle p_{k-1}^{(i)}, p_{S,k} \rangle} \tag{4.11}$$

while $f_{k|k-1}(\cdot|\zeta)$ is the *single target transition density*,$p_{S,k}(\zeta)$ is the *probability of the target's survival* and $\{(r_{\Gamma,k}^{(i)}, p_{\Gamma,k}^{(i)})\}_{i=1}^{\mathbb{M}_{\Gamma,k}}$ are the *Bernoulli components* of the birth RFS at time k. The prior multi-target density is a union of the surviving multi-Bernoulli parameter sets with the target births. Intuitively, the surviving hypothesized tracks and combined with new hypothesized tracks. As a result of this, the total number of predicted tracks is $\mathbb{M}_{k|k-1} = \mathbb{M}_{k-1} + \mathbb{M}_{\Gamma,k}$. The mean cardinality of the predicted multi-target state is [13]:

$$\bar{N}_{k|k-1} = \sum_{i=1}^{\mathbb{M}_{k-1}} r_{P,k|k-1}^{(i)} + \sum_{i=1}^{\mathbb{M}_{\Gamma,k}} r_{\Gamma,k}^{(i)} \tag{4.12}$$

The next stage of the filter is the **update** stage. At time k, given the predicted multi-target density is a multi-Bernoulli in the form of $\pi_{k|k-1} = \{(r_{k|k-1}^{(i)}, p_{k|k-1}^{(i)})\}_{i=1}^{\mathbb{M}_{k|k-1}}$, the cardinality-balanced update can be approximated as a multi-target Bernoulli in the form of:

$$\pi_k \approx \{(r_{L,k}^{(i)}, p_{L,k}^{(i)})\}_{i=1}^{\mathbb{M}_{k|k-1}} \cup \{(r_{U,k}^{*}(z), p_{U,k}^{*}(\cdot; z))\}_{z \in Z_k} \tag{4.13}$$

where

$$r_{L,k}^{(i)} = r_{k|k-1}^{(i)} \frac{1 - \langle p_{k|k-1}^{(i)}, p_{D,k} \rangle}{1 - r_{k|k-1}^{(i)} \langle p_{k|k-1}^{(i)}, p_{D,k} \rangle} \tag{4.14}$$

$$p_{L,k}^{(i)} = p_{k|k-1}^{(i)}(x)\frac{1 - p_{D,k}(x)}{1 - \langle p_{k|k-1}^{(i)}, p_{D,k}\rangle} \tag{4.15}$$

$$r_{U,k}^{*}(z) = \frac{\sum_{i=1}^{\mathbb{M}_{k|k-1}} \frac{r_{k|k-1}^{(i)}(1 - r_{k|k-1}^{(i)})\langle p_{k|k-1}^{(i)}, \psi_{k,z}\rangle}{(1 - r_{k|k-1}^{(i)}\langle p_{k|k-1}^{(i)}, p_{D,k}\rangle)^2}}{\kappa_k(z) + \sum_{i=1}^{\mathbb{M}_{k|k-1}} \frac{r_{k|k-1}^{(i)}\langle p_{k|k-1}^{(i)}, \psi_{k,z}\rangle}{1 - r_{k|k-1}^{(i)}\langle p_{k|k-1}^{(i)}, p_{D,k}\rangle}} \tag{4.16}$$

$$p_{U,k}^{*}(x; z) = \frac{\sum_{i=1}^{\mathbb{M}_{k|k-1}} \frac{r_{k|k-1}^{(i)}}{1 - r_{k|k-1}^{(i)}} p_{k|k-1}^{(i)}(x)\psi_{k,z}(x)}{\sum_{i=1}^{\mathbb{M}_{k|k-1}} \frac{r_{k|k-1}^{(i)}}{1 - r_{k|k-1}^{(i)}} \langle p_{k|k-1}^{(i)}, \psi_{k,z}\rangle} \tag{4.17}$$

$$\psi_{k,z}(x) = g(z|x)_k p_{D,k}(x) \tag{4.18}$$

and $g_k(z|x)$ is the single target measurement likelihood for z given the x at time k, $p_{D,k}(x)$ is probability of detection for the state x at time k, Z_k is the set of observed measurements at time k, and κ_k is the Poisson intensity clutter at time k. The multi-target posterior density is a union between the multi-Bernoulli RFS of the legacy tracks and the the multi-Bernoulli RFS of the measurement updated tracks [13]. The total number of posterior hypothesized tracks is a sum of the predicted tracks with with measurement corrected tracks:

$$\mathbb{M}_{k|k-1} = \mathbb{M}_{k-1} + \mathbb{M}_{\Gamma,k} \tag{4.19}$$

The mean cardinality of the posterior multi-target density is estimated as:

$$\bar{N}_k = \sum_{i=1}^{\mathbb{M}_{k|k-1}} r_{L,k}^{(i)} + \sum_{z \in Z_k}^{\mathbb{M}_{\Gamma,k}} r_{U,k}^{*}(z) \tag{4.20}$$

After CBMeMBer recursion, the states of the tracked targets are estimated. As discussed earlier, the multi-target states are characterized by their probability of existence, $r^{(i)}$, and their posterior density, $p^{(i)}$. The possibility that a hypothesized track is the true track depends on its probability of existence, $r^{(i)}$ [13]. The tracks which have higher probabilities of existence, $r^{(i)}$, are more likely true tracks. Thus, the CBMeMBer filter estimates the states by choosing the mean or mode of hypothesized tracks' posterior density with existence probabilities exceeding a certain threshold [13]. In estimating number of targets, the mode of the posterior cardinality distribution is a more

preferable choice than the mean as it is more stable [13]. Once the number of targets have been estimated from the cardinality distribution, the same number of hypothesized tracks are chosen with a descending probability of existence, $r^{(i)}$. The state of the individual tracks are estimated from the posterior density, $p^{(i)}$ of these hypothesized tracks. The general framework of the CBMeMBer is summarised in the following pseudocode:

Algorithm 3 Pseudocode for CBMeMBer

Data Initialization

1: Define the properties to be included in the multi-target state, X_k, and the multi-target measurement, Z_k.
2: Model the multi-target states and the multi-target observations as multi-Bernoulli RFS
3: Model the clutter as a Poisson RFS

Filtering

4: Predict the multi-target prior density, $\pi_{k|k-1}$, at time k based on the multi-target posterior density at time $k-1$. The multi-target prior density is a union between the surviving multi-Bernoulli parameter set with the target birth:
$\pi_{k|k-1} = \{(r^{(i)}_{P,k|k-1}, p^{(i)}_{P,k|k-1})\}_{i=1}^{\mathbb{M}_{k-1}} \cup \{(r^{(i)}_{\Gamma,k}, p^{(i)}_{\Gamma,k})\}_{i=1}^{\mathbb{M}_{\Gamma,k}}$
5: Update the predicted prior density with the received measurements. The CBMeMBer's *parameterised approximation of the multi-target Bayes posterior density* is a union between the legacy density with the measurement updated density:
$\pi_k \approx \{(r^{(i)}_{L,k}, p^{(i)}_{L,k})\}_{i=1}^{\mathbb{M}_{k|k-1}} \cup \{(r^{*}_{U,k}(z), p^{*}_{U,k}(\cdot; z))\}_{z \in Z_k}$

State Estimation

6: Determine the number of tracks based on the mean or mode of the cardinality distribution and select the same number of hypothesized tracks with the highest probability of existence, $r^{(i)}$
7: Extract the state information from the selected hypothesized tracks' posterior density

4.2.3 Sequential Monte Carlo (SMC) Implementation of the CBMeMBer

The CBMeMBer filter is to be applied in this research in order to track multiple moving speech sources. The scenario in which the CBMeMBer is to be used is the "conference room scenario" whereby multiple speakers will be moving around and will be taking turns speaking. The dynamics of these speakers are non-linear as the speakers will be moving across the room in a non-linear manner. The measurements generated by these speakers are extracted as acoustic features using BSS via signal sparsity and used for sound source localisation. TDoA of a single microphone pair which is used

to determine the location of the speakers, defines the hyperbolic surface on which the sound source is located on [76]. Hence, the measurement generated by speech sources are non-linear as well. In [13], two methods of implementation for the CBMeMBer was proposed - the SMC implementation of the CBMeMBer and the closed form linear Gaussian implementation. The linear Gaussian implementation is best suited for linear models whereas the SMC implementation of the CBMeMBer is able to accommodate non-linear source dynamics and measurement models [13]. In order to track moving speech sources which has non-linear dynamics and measurements, the SMC implementation of the CBMeMBer is more appropriate.

The SMC implementation of the CBMeMBer is similar to the particle filter used to track sound sources in [64] as the multi-target densities are presented as weighted samples formed by particles. The SMC implementation of the CBMeMBer recursively predicts and updates the particles.

At the **prediction** stage, the multi-Bernoulli posterior density, $\pi_{k-1} = \{(r_{k-1}^{(i)}, p_{k-1}^{(i)})\}_{i=1}^{\mathbb{M}_{k-1}}$, at time, $k-1$, has each of its $p_{k-1}^{(i)}$ represented by a set of weighted particles $\{w_{k-1}^{(i,j)}, x_{k-1}^{(i,j)}\}_{j=1}^{L_{k-1}^{(i)}}$. This can be explicitly expressed as:

$$p_{k-1}^{(i)}(x) = \sum_{j=1}^{L_{k-1}^{(i)}} w_{k-1}^{(i,j)} \delta_{x_{k-1}^{(i,j)}}(x) \tag{4.21}$$

with $w_{k-1}^{(i,j)}$ representing the weight and $\delta_{x_{k-1}^{(i,j)}}(x)$ representing the particles. The predicted multi-Bernoulli prior density, $\pi_{k|k-1}$ is calculated as:

$$\pi_{k|k-1} = \{(r_{P,k|k-1}^{(i)}, p_{P,k|k-1}^{(i)})\}_{i=1}^{\mathbb{M}_{k-1}} \cup \{(r_{\Gamma,k}^{(i)}, p_{\Gamma,k}^{(i)})\}_{i=1}^{\mathbb{M}_{\Gamma,k}} \tag{4.22}$$

with the particle representation of $r_{P,k|k-1}^{(i)}$, $p_{P,k|k-1}^{(i)}$, $r_{\Gamma,k}^{(i)}$ and $p_{\Gamma,k}^{(i)}$ being:

$$r_{P,k|k-1}^{(i)} = r_{k-1}^{(i)} \sum_{j=1}^{L_{k-1}^{(i)}} w_{k-1}^{(i,j)} p_{S,k}(x_{k-1}^{(i,j)}) \tag{4.23}$$

$$p_{P,k|k-1}^{(i)}(x) = \sum_{j=1}^{L_{k-1}^{(i)}} \tilde{w}_{P,k|k-1}^{(i,j)} \delta_{x_{P,k|k-1}^{(i,j)}}(x) \tag{4.24}$$

$$r_{\Gamma,k}^{(i)} = \text{birth model's parameter} \tag{4.25}$$

$$p_{\Gamma,k}^{(i)}(x) = \sum_{j=1}^{L_{\Gamma,k}^{(i)}} \tilde{w}_{\Gamma,k}^{(i,j)} \delta_{x_{\Gamma,k}^{(i,j)}}(x) \tag{4.26}$$

whereby the variables, $x_{P,k|k-1}^{(i,j)}$, $w_{P,k|k-1}^{(i,j)}$, $\tilde{w}_{P,k|k-1}^{(i,j)}$, $x_{\Gamma,k}^{(i,j)}$, $w_{\Gamma,k}^{(i,j)}$, and $\tilde{w}_{\Gamma,k}^{(i,j)}$ are defined as [13]:

$$x_{P,k|k-1}^{(i,j)} \sim q_k^{(i)}(\cdot|x_{k-1}, Z_k),\ j = 1, \cdots, L_{k-1}^{(i)} \tag{4.27}$$

$$w_{P,k|k-1}^{(i,j)} = \frac{w_{k-1}^{(i,j)} f_{k|k-1}(x_{P,k|k-1}^{(i,j)}|x_{k-1}^{(i,j)}) p_{S,k}(x_{k-1}^{(i,j)})}{q_k^{(i)}(x_{P,k|k-1}^{(i,j)}|x_{k-1}, Z_k)} \tag{4.28}$$

$$\tilde{w}_{P,k|k-1}^{(i,j)} = \frac{w_{P,k|k-1}^{(i,j)}}{\sum_{j=1}^{L_{k-1}^{(i)}} w_{P,k|k-1}^{(i,j)}} \tag{4.29}$$

$$x_{\Gamma,k}^{(i,j)} \sim b_k^{(i)}(\cdot|Z_k)\ j = 1, \cdots, L_{\Gamma,k}^{(i)} \tag{4.30}$$

$$w_{\Gamma,k}^{(i,j)} = \frac{p_{\Gamma,k}(x_{\Gamma,k}^{(i,j)})}{b_k^{(i)}(x_{\Gamma,k}^{(i,j)}|Z_k)} \tag{4.31}$$

$$\tilde{w}_{\Gamma,k}^{(i,j)} = \frac{w_{\Gamma,k}^{(i,j)}}{\sum_{j=1}^{L_{\Gamma,k}^{(i)}} w_{\Gamma,k}^{(i,j)}} \tag{4.32}$$

on the condition that the importance densities $q_k^{(i)}(\cdot|x_{k-1}, Z_k)$ is given such that $support(p_k^{(i)}) \subseteq support(q_k^{(i)})$ and $b_k^{(i)}(\cdot|Z_k)$ such that $support(p_\Gamma^{(i)}) \subseteq support(b_k^{(i)})$ [13].

The **update** stage comes after the prediction stage. The multi-Bernoulli posterior density, π_k, at time, k, can be calculated on the condition that the predicted multi-Bernoulli prior density , $\pi_{k|k-1} = \{(r_{k|k-1}^{(i)}, p_{k|k-1}^{(i)})\}_{i=1}^{\mathbb{M}_{k|k-1}}$, is known and each $p_{k-1}^{(i)}, i = 1, \cdots, \mathbb{M}_{k|k-1}$ is represented by a weighted set of particles, $\{w_{k|k-1}^{(i,j)}, x_{k|k-1}^{(i,j)}\}_{j=1}^{L_{k|k-1}^{(i)}}$ i.e.:

$$p_{k|k-1}^{(i)}(x) = \sum_{j=1}^{L_{k|k-1}^{(i)}} w_{k|k-1}^{(i,j)} \delta_{x_{k|k-1}^{(i,j)}}(x) \tag{4.33}$$

The particle representation of CBMeMBer's *parameterised approximation of the multi-target Bayes posterior density*, $\pi_k \approx \{(r_{L,k}^{(i)}, p_{L,k}^{(i)})\}_{i=1}^{\mathbb{M}_{k|k-1}} \cup \{(r_{U,k}^{*}(z), p_{U,k}^{*}(\cdot; z))\}_{z \in Z_k}$ can be calculated as [13]:

$$r_{L,k}^{(i)} = r_{k|k-1}^{(i)} \frac{1 - \varrho_{L,k}^{(i)}}{1 - r_{k|k-1}^{(i)} \varrho_{L,k}^{(i)}} \tag{4.34}$$

$$p_{L,k}^{(i)}(x) = \sum_{j=1}^{L_{k|k-1}^{(i)}} \tilde{w}_{L,k}^{(i,j)} \delta_{x_{k|k-1}^{(i,j)}}(x) \tag{4.35}$$

$$r^*_{U,k}(z) = \frac{\sum_{i=1}^{\mathbb{M}_{k|k-1}} \frac{r^{(i)}_{k|k-1}(1-r^{(i)}_{k|k-1})\varrho^{(i)}_{U,k}(z)}{(1-r^{(i)}_{k|k-1}\varrho^{(i)}_{L,k})^2}}{\kappa_k(z) + \sum_{i=1}^{\mathbb{M}_{k|k-1}} \frac{r^{(i)}_{k|k-1}\varrho^{(i)}_{U,k}(z)}{1-r^{(i)}_{k|k-1}\varrho^{(i)}_{L,k}}} \tag{4.36}$$

$$p^*_{U,k}(x;z) = \sum_{i=1}^{\mathbb{M}_{k|k-1}} \sum_{j=1}^{L^{(i)}_{k|k-1}} \tilde{w}^{*(i,j)}_{U,k}(z)\delta_{x^{(i,j)}_{k|k-1}}(x) \tag{4.37}$$

whereby the variables $\varrho^{(i)}_{L,k}$, $w^{(i,j)}_{L,k}$, $\tilde{w}^{(i,j)}_{L,k}$, $\varrho^{(i)}_{U,k}(z)$, $w^{*(i,j)}_{U,k}(z)$, and $\tilde{w}^{*(i,j)}_{U,k}(z)$ are defined as [13]:

$$\varrho^{(i)}_{L,k} = \sum_{j=1}^{L^{(i)}_{k|k-1}} w^{(i,j)}_{k|k-1} p_{D,k}(x^{(i,j)}_{k|k-1}) \tag{4.38}$$

$$w^{(i,j)}_{L,k} = w^{(i,j)}_{k|k-1}(1 - p_{D,k}(x^{(i,j)}_{k|k-1})) \tag{4.39}$$

$$\tilde{w}^{(i,j)}_{L,k} = \frac{w^{(i,j)}_{L,k}}{\sum_{j=1}^{L^{(i)}_{k|k-1}} w^{(i,j)}_{L,k}} \tag{4.40}$$

$$\varrho^{(i)}_{U,k}(z) = \sum_{j=1}^{L^{(i)}_{k|k-1}} w^{(i,j)}_{k|k-1}\psi_{k,z}(x^{(i,j)}_{k|k-1}) \tag{4.41}$$

$$w^{*(i,j)}_{U,k}(z) = w^{(i,j)}_{k|k-1}\frac{r^{(i)}_{k|k-1}}{1 - r^{(i)}_{k|k-1}}\psi_{k,z}(x^{(i,j)}_{k|k-1}) \tag{4.42}$$

$$\tilde{w}^{*(i,j)}_{U,k}(z) = \frac{w^{*(i,j)}_{U,k}(z)}{\sum_{i=1}^{\mathbb{M}_{k|k-1}} \sum_{j=1}^{L^{(i)}_{k|k-1}} w^{*(i,j)}_{U,k}(z)} \tag{4.43}$$

The SMC implementation of the CBMeMBer suffers from the same *degeneracy problem* that plagues SMC algorithms and **resampling** is required to reduce the effect this problem [13]. In SMC algorithms, the *degeneracy problem* occurs when most of the weights are close to zero and only a select few weights are updated [71]. By having the resampling process, particles with low weights are eliminated from the sampling pool while particles with high weights are multiplied so the computational effort can be focused on regions with higher probability of target existence [111]. As a result of this, the particle distribution will be focused on zones in the room where the sound sources are more likely to exist. There are three main types of resampling scheme - Systemic Resampling, Residual Resampling and Multinomial Resampling. The computational

complexity and the Monte Carlo error will be affected by the resampling technique used [112]. As a result of this, the resampling technique used in the proposed solution is the Systemic Resampling technique as it is easy to implement and it outperforms most resampling schemes in most scenarios [111] although there is no certainty this in general [112]. Resampling of the particles for each of the hypothesized track is carried out after the prediction step. As the complexity scales linearly, computational load is an issue when the number of targets increases.

As the number of hypothesized tracks increases due to target births, the number of particles required to represent the multi-target posterior density also increases. Pruning is required in order to reduce the computational load of the algorithm [13]. Pruning is the process of removing hypothesized tracks with existence probability below a certain threshold, p_{thres}. The remaining hypothesized tracks are allocated the number of particles proportional to its probability of existence as it is desirable to allocate the number of particles in the track density in proportion to the number of expected targets [13]. An example of this is sampling of $L_{\Gamma,k}^{(i)} = r_{\Gamma,k}^{(i)} L_{\max}$ number of particles during the prediction step and resampling the particles according to $L_k^{(i)} = r_k^{(i)} L_{\max}$ for each of the updated hypothesized tracks during the resampling step. In order to control the computational load of the algorithm, the number of particles allocated to each track is given a lower bound limit of $L_{\min}$ and an upper bound limit of $L_{\max}$

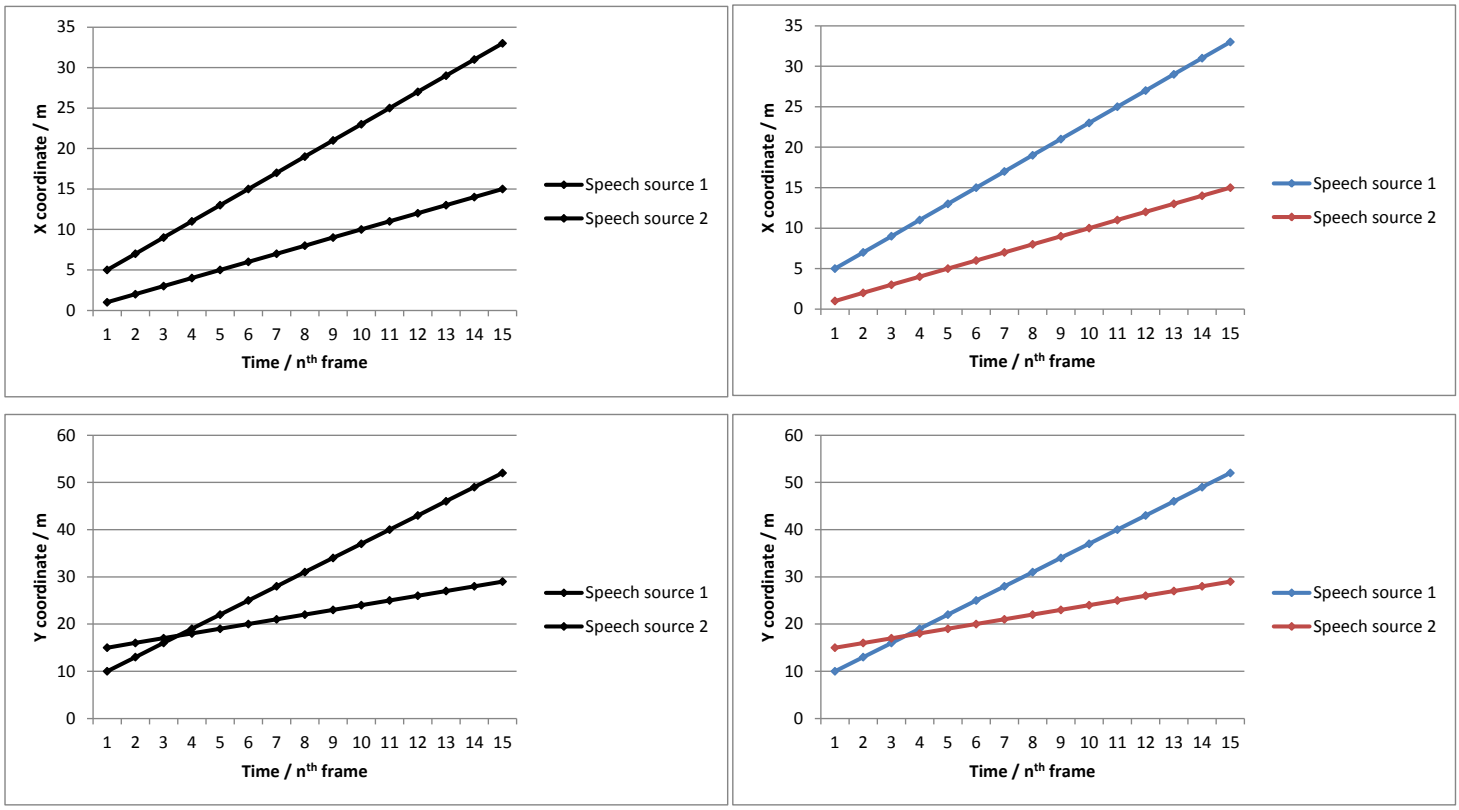
The final step in the SMC recursion is the multi-target state estimation. The implementation of the multi-target state estimation technique is fairly straight forward as the number of targets tracked are first estimated from the cardinality's mean or mode. Despite the fact that mean of the cardinality is simpler to implement, the mode of the cardinality is used in the proposed solution as the mode of the cardinality is more stable [13]. An equal number of hypothesized tracks with the highest number of existence probability are then chosen for multi-target state estimation. The individual states are obtained by calculating the means of the corresponding posterior density associated with the selected hypothesized tracks. The multi-target state estimation of the SMC-CBMeMBer has a lower computational complexity than the SMC implementation of the PHD or the CPHD filter. In the SMC implementation of PHD or the CPHD filter, multi-target state estimation is complicated - the number of states are first estimated from the cardinality's mean or mode and corresponding clusters are formed from the

particles of the intensity function with the centers of these cluster being the multi-target state estimates. As a result of clustering, the SMC implementation of PHD or CPHD has a few associated problems. The first of this is the associated high computational complexity of clustering which does not scale well with the increase in the number of targets tracked. Moreover, the estimated number of targets has to match the natural number of clusters; else the multi-target state estimates will not be accurate [13].

4.3 Application of CBMeMBer in Speech Source Tracking

In this research, the CBMeMBer filter was chosen to solve the problem of tracking multiple moving speech sources in a "conference room scenario". In speech source tracking, the function of the CBMeMBer filter is to filter the true acoustic measurements from the acoustic features extracted by the BSS framework and estimate the number of speech sources in the speech source mixture. The true acoustic measurements are used to localise and track the speech sources. CBMeMBer is an appropriate technique to be used in speech source tracking due to its various strengths compared to other multi-target tracking techniques. First of all, it is an RFS technique which is mathematical principled as it uses Finite Set Statistics (FISST) [12] to extend and cast the single target tracking problem to a multi-target tracking problem in a Bayesian tracking framework. The CBMeMBer filter has formal target birth and death models to represent the spontaneous appearance and disappearance of the speech sources. The CBMeMBer is also a mathematically tractable technique as it propagates the multi-Bernoulli approximate of the multi-target posterior density instead of the full multi-target posterior density. Apart from that, the CBMeMBer filter has also been shown to have better performance than the PHD filter despite have similar computational complexity in terms of target scaling [13].

The main limitation of the CBMeMBer in the application of speech source tracking is the incapability of distinguishing the different speech sources and their trajectories. In the CBMeMBer filter, labels are not propagated as a parameter in the multi-Bernoulli density to reduce the computer complexity of the CBMeMBer filter. Data association is implicitly available as the measurements are matched to the individual



(a) Estimated tracks with no labels

(b) Estimated tracks with labels

Figure 4.2: Example of estimated tracks without labels and with labels

posterior densities and propagated but there is no explicit labelling method in order to distinguish the different trajectories of the targets. Although this is a boon which allows the CBMeMBer to be more tractable in other tracking applications, it is a bane in speech source tracking as the filtered acoustic features need to be matched with their associated speech sources. When the speech sources are moving, the acoustic features of the speech sources vary across the time frames. The identity information is required in order to construct the unique TF masks from these acoustic features which belong to the same sound source. Figure 4.2 is an example illustrating the differences between estimated tracks without labels and estimated tracks with labels. An example of the direct output from CBMeMBer is shown in sub-figure 4.2a while an output with the addition of labels is shown in sub-figure 4.2b.

Visually, humans can easily tell the tracks apart. Without such visual cues, labels or identity information is required by the machine to match the tracks with their respective sound sources. The identity information is crucial for the algorithm to tell the tracks apart yet the identity information is not incorporated into the state vector of the

speech sources in CBMeMBer in order to lower the computational complexity of the algorithm.

In order to overcome the limitation of CBMeMBer which lacks an explicit data association component, [13] has presented an extension to allow for track propagation. The proposed extension used a track table, $\mathcal{T}$, consisting of the triplets - probability of existence, $r^{(i)}$, probability density, $p^{(i)}$, and label, $l^{(i)}$, to identify the different tracks [13]. The label, $l^{(i)}$, is assigned to each of the Bernoulli component in the multi-Bernoulli density.

Using the extension, the track table was propagated with each recursion. During the **prediction** stage, the labels, $l^{(i)}$, are propagated along with the existence probabilities, $r^{(i)}$, and the probability densities, $p^{(i)}$. The prior track table is predicted in place of the multi-Bernoulli density. The prior track table can be calculated as:

$$\mathcal{T}_{k|k-1} = \{(l^{(i)}_{P,k|k-1}, r^{(i)}_{P,k|k-1}, p^{(i)}_{P,k|k-1})\}_{i=1}^{M_{k-1}} \cup \{(l^{(i)}_{\Gamma,k}, r^{(i)}_{\Gamma,k}, p^{(i)}_{\Gamma,k})\}_{i=1}^{M_{\Gamma,k}} \tag{4.44}$$

with $l^{(i)}_{P,k|k-1}$ inheriting the label, $l^{(i)}_{k-1}$ and $l^{(i)}_{\Gamma,k}$ denoting a new label resulting from the birth of a new target. $r^{(i)}_{P,k|k-1}$, $p^{(i)}_{P,k|k-1}$, $r^{(i)}_{\Gamma,k}$ and $p^{(i)}_{\Gamma,k}$ retains the same definition as their prior multi-Bernoulli counterpart in equation (4.9). In the **update** stage, the track table also replaces the multi-Bernoulli posterior density. Given the predicted track table at time, k, is, $\mathcal{T}_{k|k-1} = \{(l^{(i)}_{k|k-1}, r^{(i)}_{k|k-1}, p^{(i)}_{k|k-1})\}_{i=1}^{M_{k|k-1}}$, the updated track table can be calculated as:

$$\mathcal{T}_k \approx \{(l^{(i)}_{L,k}, r^{(i)}_{L,k}, p^{(i)}_{L,k})\}_{i=1}^{M_{k|k-1}} \cup \{(l_{U,k}(z), r^{*}_{U,k}(z), p^{*}_{U,k}(\cdot; z))\}_{z \in Z_k} \tag{4.45}$$

with $l^{(i)}_{L,k}$ being $l^{(i)}_{k|k-1}$ and $l_{U,k}(z) = l^{(n)}_{L,k}$ whereby:

$$n = \arg\max_{i} \frac{r^{(i)}_{k|k-1}(1 - r^{(i)}_{k|k-1})\langle p^{(i)}_{k|k-1}, \psi_{k,z}\rangle}{(1 - r^{(i)}_{k|k-1}\langle p^{(i)}_{k|k-1}, p_{D,k}\rangle)^2} \tag{4.46}$$

while $r^{(i)}_{L,k}$, $p^{(i)}_{L,k}$, $r^{*}_{U,k}(z)$ and $p^{*}_{U,k}(\cdot; z))$ remain the same as their multi-Bernoulli counterparts in equation (4.13).

The labelling method proposed in [13] is an intuitive method to address the problem of target and trajectory identification but it has its limitations. One of the limitations of this labelling extension is the poor performance of it when two targets are close together [13]. If two targets are close to each other and they are merged as a single target in one frame, and then separated in a different frame, one of the targets will

be assigned a new label instead of retaining its label prior to the merging [113]. This will be a problem in speech source tracking when the speakers are close together. Furthermore, this problem may also arise when two speech sources have the same TDoA despite not being physically close together. This is because the TDoA from a single pair of microphones defines the hyperbolic surface on which the speech source may lie [76]. In speech source tracking, the speakers take turns speaking so there will be silence periods in the speech sources. With the labelling extension proposed in [13], the speech sources will be assigned a new identity. This labelling extension is unable to resolve the identity issues in the aforementioned acoustic scenarios because this method only associates labels directly to the past time step labels [114].

Despite this limitation, it is more beneficial to have this extension rather than incorporating labels as part of the states propagated in order to maintain a lower computational complexity. By incorporating the labels as part of the states in the Bayesian recursion, the computational complexity will increase due to the permutations involved in associating the labels with the measurements and hypothesized tracks. As this research aims to provide an online solution, a computationally low yet robust method which assigns the labels for the different track trajectories based on their dynamics will be required. To overcome the identity association problem that occurs in sound source tracking, a labelling extension which is capable associating the CBMeMBer estimated tracks with their respective speech sources is proposed in Chapter 5. The availability of the identity information is crucial as the capability of associating the estimated states with their respective speech sources enables the speech source mixture to separated by TF masks.

4.4 Summary

In summary, Chapter 4 has addressed the problem of jointly estimating the number and state of the speech sources after the acoustic features have been extracted using the BSS framework. The problem of joint estimation on the number and the state of the speech sources is a multi-target estimation problem and RFS multi-target framework which falls under FISST is a mathematically principled way of solving this problem [12]. The RFS method chosen for this research is the SMC-CBMeMBer. Compared to PHD

or CPHD, the CBMeMBer is preferred technique as it has a computational complexity which scales linearly with the targets and the measurements and in certain scenarios, it outperforms both the PHD and CPHD filters [13]. The SMC implementation of CBMeMBer selected as it is designed to accommodate non-linear scenarios [13] and the dynamics and measurement models for speech sources are non-linear.

In CBMeMBer, the multi-target posterior density is modelled as a multi-Bernoulli RFS while the clutter is modelled as a Poisson RFS. The two stages in CBMeMBer recursion are the prediction and update stages. The prior multi-target density is estimated as $\pi_{k|k-1} = \{(r^{(i)}_{P,k|k-1}, p^{(i)}_{P,k|k-1})\}_{i=1}^{M_{k-1}} \cup \{(r^{(i)}_{\Gamma,k}, p^{(i)}_{\Gamma,k})\}_{i=1}^{M_{\Gamma,k}}$, during the **prediction**. In the **update** stage, CBMeMBer propagates the approximated the multi-target Bernoulli parameter set instead of the full Bayes recursion as the full Bayes recursion is computational expensive to propagate [13]. The CBMeMBer's *parameterised approximation of the multi-target Bayes posterior density* is estimated as $\pi_k \approx \{(r^{(i)}_{L,k}, p^{(i)}_{L,k})\}_{i=1}^{M_{k|k-1}} \cup \{(r^{*}_{U,k}(z), p^{*}_{U,k}(\cdot; z))\}_{z \in Z_k}$. The multi-target state estimation is performed after the prediction stage. The number of targets can be estimated from the mean or mode of the cardinality distribution. In this research, the mode was chosen as it is more stable than the mean. The multi-target states are estimated from the probability densities of the corresponding number of hypothesized tracks with the highest probability of existence.

The main limitation of the CBMeMBer in the application of acoustic tracking is the lack of a labelling technique to distinguish the different trajectories of the tracked speech sources. An extension that made use of track table, $\mathcal{T}$, was proposed in [13] but the performance of this heuristic method was expected to be poor when it is applied to speech source tracking. Although this labelling technique was not optimal, the idea of not inherently propagating the labels helps to reduce the computational load required to track the multiple targets. Furthermore, the extension also shows the viability of CBMeMBer filter to be combined with an external identity association algorithm in order to assign labels to the track. Thus, a more robust labelling extension is proposed in Chapter 5 to overcome the problem of identity association.

Chapter 5

Speech Source Localisation, Tracking and Separation

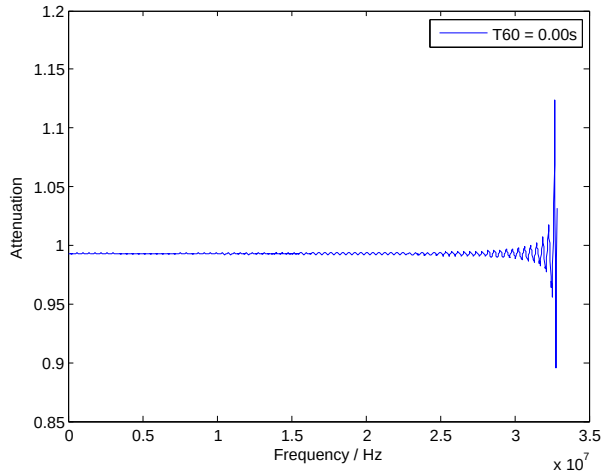
There is nothing wrong with imitating someone,
as long as it is to find who you really are.
– Tendou Souji

5.1 Overview

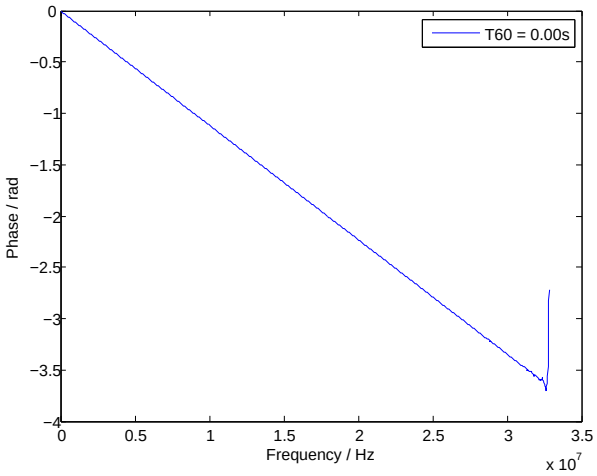
The novelty of this research is to apply the concepts of BSS via signal sparsity and combine these concepts with a CBMeMBer filter along with track management algorithm to solve the speech source separation problem in the "conference room scenario". The speech source separation problem in the "conference room scenario" is a challenging problem to solve due to factors such as the lack of *a priori* information on the number of speech sources and the movement of these speech sources. Furthermore, the problem is exacerbated with the presence of noise and reverberation in a room. Hence, the complex "conference room scenario" speech source separation problem is broken down into three smaller problems - acoustic feature extraction from source mixture, source number and state estimation based on the extracted acoustic features and the final problem of source separation. These three main aspects of the "conference room scenario" speech source separation problem has to be solved in order to localise, track and separate the unknown number of moving speech sources.

The problem of acoustic feature extraction and estimation for non-moving sound

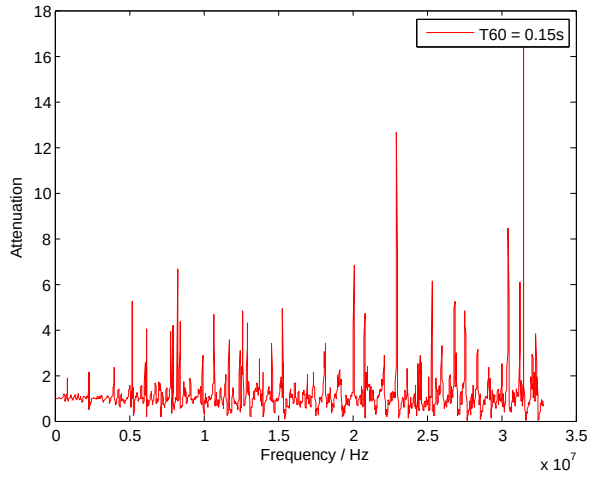
sources is a very difficult problem on its own. An example of this is illustrated in figures 5.1.



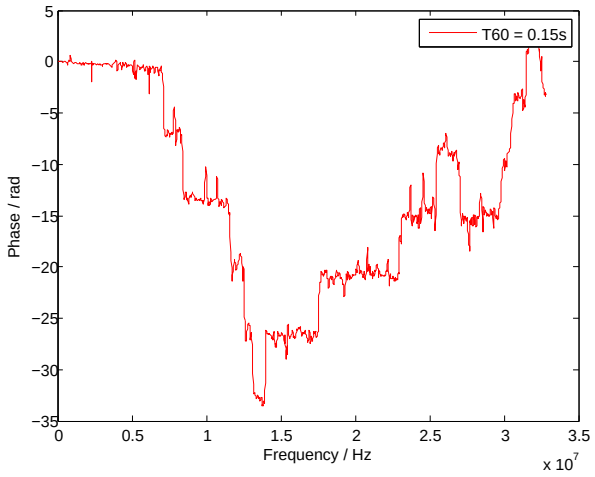
(a) Attenuation estimation in room reverberation of T60 = 0.00s



(b) Phase estimation in room reverberation of T60 = 0.00s



(c) Attenuation estimation in room reverberation of T60 = 0.15s



(d) Phase estimation in room reverberation of T60 = 0.15s

Figure 5.1: Attenuation and phase estimation in different room reverberation

Both the attenuation and delay (represented in phase form) are corresponding to the true position only in an ideal scenario where there is no noise or reverberation in the room. When the room reverberation is increased, the measured attenuation and delay fluctuates. The arduous task of acoustic features extraction and estimation is further compounded when the speech sources are moving. When the speech sources are moving, these acoustic features will change relative to the position of the sources.

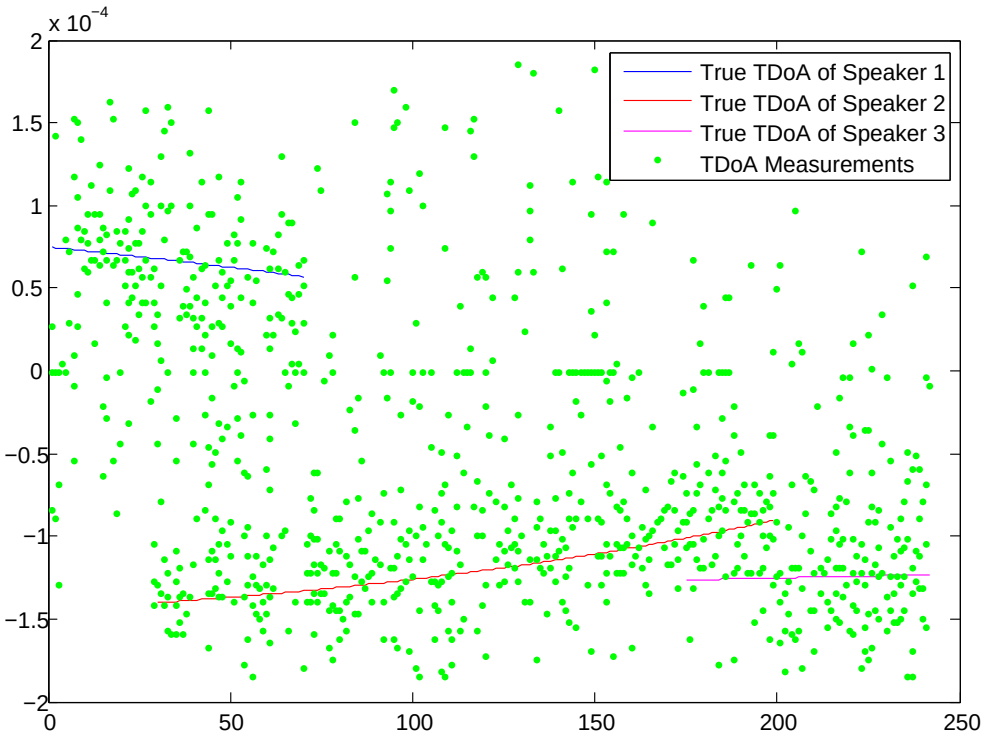


Figure 5.2: Measurements of TDoA extracted by BSS via Signal Sparsity

The problem of acoustic features extraction from source mixture through the use of BSS via signal sparsity and its limitations in the application of moving speech sources have been discussed in Chapter 3.

After the acoustic features have been extracted, the state and the number of the speech sources are to be jointly estimated by the CBMeMBer filter. The problem of jointly estimating the state and the number of speech sources based on the acoustic feature measurements is also another difficult problem to solve. The CBMeMBer has to predict the movement of the speech sources based on the dynamics model of a typical human speaker and confirm the likelihood of the predictions based on the measurement updates. The speech sources may or may not generate measurements and these measurements are corrupted by clutter. An example of the TDoA measurements based on the extracted acoustic features is shown in figure 5.2. As illustrated in figure 5.2, the task of filtering out the true measurements generated by the speech sources is challenging as the true TDoA generated by the speech sources are obscured by clutter and it is barely discernible using human vision. The background information and the principles applied by the CBMeMBer to jointly estimate the number and state of the tracked sound sources were reviewed in Chapter 4.

With the acoustic features extracted and the states as well as the number of the speech sources estimated, the sound source separation problem in the "conference room scenario" Chapter 5 will focus on the main contribution of this research which is the integration of the BSS via signal sparsity framework and CBMeMBer along with a track management extension to solve the speech source separation problem in "conference room scenario". The problem of moving speech source separation can be achieved by applying the principle of signal sparsity to extract and separate the acoustic features and to track the speech sources with CBMeMBer. In order to integrate the BSS framework and CBMeMBer filter algorithms to perform online localisation, tracking and separation of moving speech sources, adaptations have to be made to both algorithms.

The main features of both BSS via signal sparsity and CBMeMBer in the "conference room scenario" drive the motivation for these adaptations. The BSS via signal sparsity algorithm can be adapted to provide instantaneous estimates of the speech sources' acoustic features while the CBMeMBer filter can be adapted to use highly uncertain data as is the case for acoustic speech sources. Once the adaptations are made, both concepts of BSS via signal sparsity and the CBMeMBer can be integrated via the multi-target Bayesian framework. The acoustic features extracted using the BSS framework will be used as measurements for the CBMeMBer. The CBMeMBer rely on the extracted acoustic features to perform a joint estimation on the number and the state of the sound sources. The output of the CBMeMBer can then be post processed by a track management system and used to guide the construction of time-frequency masks. Sound source separation can then be achieved by applying these TF masks onto the sound mixtures.

5.2 Adaptation of BSS via Signal Sparsity for Online Speech Source Separation

The sound source separation problem of the more complex "conference room scenario" requires the ability to extract audio cues from an unknown number of multiple moving speech sources and the original BSS via signal sparsity algorithm is incapable

of extracting these audio cues. This is because the original algorithm relies on the assumptions that the number of sound sources are known *a priori* and the sound sources are not moving. In order to overcome the limitations of the BSS via signal sparsity framework in extracting audio cues from moving speech sources, certain techniques used in the original algorithm have to be adapted. This means that the clustering needs to be replaced by a more flexible framework of operation. The basis for this framework is the CBMeMBer filter. The CBMeMBer filter gives an estimate of the number and positions of the sources for every frame which means that the important clustering parameters in BSS can be replaced accordingly. This will be discussed in detail in this section.

5.2.1 Recursive Acoustic Feature Extraction

The first adaptation required by the original BSS via signal sparsity algorithm in order to the extract acoustic features from the moving speech sources is to process the speech source mixture recursively on a frame-by-frame basis. In the original BSS via signal sparsity, the sound sources were assumed to be non-moving so the acoustic transfer function or TF ratio as it is known in [4] is calculated according to equation (3.4)

$$\mathcal{H}_{21} = \frac{Y_2(\tau,\omega)}{Y_1(\tau,\omega)} \tag{5.1}$$

whereby Y_1 and Y_2 are Short-Time Fourier Transform (STFT) of the signals received by the microphone pair. Equation (3.4) is only valid for non-stationary and non-moving sound sources. The relative attenuation, a_n, and the relative delay, δ_n can be extracted by exploiting the sparsity of the signals

$$\mathcal{H}_{21} = \frac{a_n e^{-j\omega\delta_n} S_n(\tau,\omega)}{S_n(\tau,\omega)} = a_n e^{-j\omega\delta_n}. \tag{5.2}$$

The values of a_n and δ_n remains the same for each sound source throughout the whole signal when the sound sources are non-stationary and non-moving. When the sound sources are moving, the transfer function changes as well so the original acoustic model has to be modified to adapt to these changes for online implementation. The relative attenuation, a_n and the relative delay δ_n are now dependent on the position of the sound sources relative to the microphone. The shift in the delay and attenuation when the sound sources are moving will affect the system's capability to separate the sound

signals. A better TF ratio suggested in [115] and [116] uses the cross power spectral density of the signal with its auto power spectral density would be a better choice to obtain the acoustic features for moving speech sources

$$\mathcal{H}_{21} = \frac{Y_2(\tau,\omega)Y_1^*(\tau,\omega)}{Y_1(\tau,\omega)Y_1^*(\tau,\omega)} = a_{mn,x_{k,n}} e^{-j\omega\delta_{mn,x_{k,n}}} \tag{5.3}$$

whereby $a_{mn,x_{k,n}}$ and $\delta_{mn,x_{t,n}}$ are still the relative attenuation and relative delay between the sound source and microphones at time k. By altering the original algorithm to process the input sound mixture in a recursive manner, not only can the acoustic features from moving sound sources be extracted, the effects of spurious peaks resulting from noise and reverberation can also be mitigated. As shown in [64], the effects of the spurious peaks on the accuracy of the acoustic features extraction can be mitigated by introducing a spatio-temporal relationship for the speech sources using a dynamics model. The true speech sources will move according to a predictable trajectory from frame to frame but the peaks due to noise and reverberation will be spurious.

In the original BSS algorithm, the whole duration of the sound mixture was processed by the power-weighted histogram. With the change in the acoustic transfer function to process the sound mixture recursively as shown in equation (5.3), a power-weighted histogram is created for each time frame of the STFTed signals. This allows the acoustic features of the sound mixture to be extracted in an iterative manner. The size of the time frame chosen for the analysis is crucial [37]. The time period of the window has to be short enough that the moving speech sources are assumed to be pseudo-stationary for each time frame [117] [116] yet not so short that it does not capture statistical data from the audio mixture. On the other hand, if the window size used for the analysis is too long, the moving speech sources will no longer be pseudo-stationary within the time frame as the changes in the targets' positions are reflected in the change of their acoustic features.

5.2.2 Acoustic Feature Selection

In the original BSS via signal sparsity algorithm, both the acoustic features - attenuation and delay were extracted and estimated. In this research, it is proposed that only the delay be used as the measurement for CBMeMBer due to the inaccuracy of

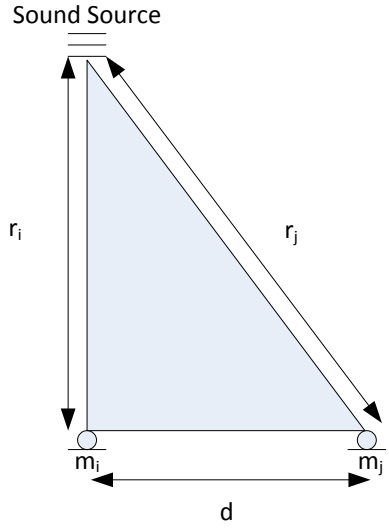


Figure 5.3: Relationship between Sound Source and Acoustic Features

the estimated attenuation. The relationship between the acoustic features and the relative positions of the sound source and microphones is illustrated in figure 5.3. The waveform of the tracked speech sources is assumed to be far field planar waves. As shown in figure 5.3, the relative attenuation is estimated by taking the ratio of the distance from the sound source to the microphone pair. As the distance from the speech source to the microphone pair, r_i, and r_j is relatively larger than the distance between the microphone pair, d, the difference between the distance between microphone pair and the sound source, r_i, and r_j becomes negligible. Hence, the estimated relative attenuation will be unable to provide useful data in terms of localisation. As a result of this, only the relative delay is transformed into the TDoA between microphone pairs and used as observations in the CBMeMBer. The centre of the peaks were determined manually in the original BSS via signal sparsity algorithm [36] but a peak finder is used in the proposed solution to obtain the measurements for the CBMeMBer filter in order to automate the process of determining the acoustic features from the peaks of the power-weighted histograms. Two parameters were used as thresholds in the peak finder with one limiting the minimum distance between two histogram peaks while the other limiting the minimum TF-weight of the histogram. The peak values of power-weighted histograms which are above the minimum TF-weight threshold and not within the minimum distance threshold are selected.

As discussed in Section 3.3, conventional BSS via signal sparsity algorithms use

clustering techniques such as power-weighted histogram, k-mean clustering or fuzzy-c means to cluster the instantaneous attenuation and delay in order to estimate the true attenuation and delay of the respective sound sources. When the speech sources are moving, the attenuation and delay changes relative to the positions of the speech sources and microphones. Hence, these clustering methods are not feasible to estimate the acoustic features of the moving speech sources. In this research, the power-weighted histogram is only used to extract the acoustic features for a time frame within which the speech sources are assumed to be pseudo-stationary. The estimation of the true acoustic features and the association of these features with their respective speech sources is processed by the CBMeMBer technique with a track management extension. By using the instantaneous acoustic features as measurements, the CBMeMBer filter estimates the number and the state of the speech sources and produces hypothesized tracks. Using these estimated tracks, a track management associates the hypothesized tracks with their respective speech sources. With the state and identity information of the speech sources known, the true acoustic sources of the relative to the position of the speech sources and the microphones can be calculated. Further details of the track management extension will be discussed in Section 5.4.1 while the details on the acoustic features estimation will be explained in Section 5.4.2.

5.2.3 TF-weight Parameter Tuning

As stated in Section 3.3, BSS via signal sparsity relies on the *W-disjoint orthogonality* assumption to achieve speech source separation. Due to this assumption, the algorithm requires the direct path components of the impulse response from the sound sources to be strong and the multipaths resulting from noise and reverberation to be weak. When this assumption is not fully met, the sound paths from the noise and reverberation will show up spurious peaks in the power-weighted histograms. In the original algorithm, the parameters p and q of the time-frequency weights, $|Y_1(\tau,\omega)Y_2(\tau,\omega)|^p\omega^q$. As shown in Section 3.2.2, the parameters p and q are advised to be tuned to $p = 2$ and $q = 2$ in order to suppress the spurious peaks resulting from noise and reverberation. The downside of having this parameter set is the suppression of true sound sources when one of the sound sources is dominant. Accounting for scenarios when one of the sound source is dominant, p is tuned to $p = 0.5$ but the spurious peaks resulting from noise

and reverberation will be enhanced too. This will not be a problem as the CBMeMBer is used to filter out the true measurements generated by the speech source from the ones generated by noise and reverberation.

5.2.4 Increase in the Number of Microphone Pairs

The BSS via signal sparsity algorithm used in this research is inspired by DUET. One of the aims of this research is to localise the speech sources in their Cartesian coordinates which is two dimensional instead of just the DoA which is just one dimensional. However, the original DUET only processes the input from a single pair of microphones. With just a single pair of sensors, only the DoA of speech sources can be determined. There is not enough information available in order to localise the Cartesian coordinates of the speech sources. Moreover, it was shown in [118] that a single microphone pair will only have unique TDoA identifier in a small region in front of the microphone pair and the error region is larger at the sides of the microphone pair. Hence, multiple microphone pairs which are spread out across the room are required to provide a more accurate tracking of the multiple sound sources. The BSS via signal sparsity algorithm used in this research processes the input signals from the multiple microphone pairs iteratively. The acoustic feature extraction process is repeated for each pair of the microphone pairs.

5.3 Adaptation of CBMeMBer for Speech Source Tracking

As mentioned earlier in Section 4.2, the CBMeMBer operates in the multi-target state space and multi-target observable space. In order to integrate the BSS via signal sparsity framework with the CBMeMBer algorithm, the single target dynamics model and the observation model has to be tailored to the scenario for moving speech sources. These adaptations include the selection of appropriate dynamics and observation model as well as a modification to the likelihood function in order to fuse the different data available from multiple sensors provided by the BSS via signal sparsity algorithm.

5.3.1 Dynamics Model

In section 3.3, it was explained that the conventional BSS via signal sparsity limitation of only separating non-stationary non-moving sound sources can be overcome by analysing the sound mixture signal on a frame by frame basis. A frame-by-frame acoustic feature extraction suffers from the spurious peaks due to noise and reverberation. In [64], it was shown that the effects of these spurious peaks can be mitigated if a spatio-temporal relationship can be established for the main sound sources. The true moving sound sources will follow specific trajectories whereas the spurious sound sources do not move according to any specific patterns. The trajectory of the sound source from frame-to-frame is described by the transition model which is also known as the dynamics model.

The choice of dynamics model to represent the trajectory of the sources can affect the tracking accuracy [74]. There are several popular dynamics models such as the constant velocity model, the constant acceleration model and also the multiple mixed model [119, p.200]. These models are suitable for vehicle tracking as vehicle movements are more linear. In speech tracking, the movements of the speakers are more random and non-linear. The effect of a few different dynamics models such as the coordinate-uncoupled model, curvilinear model, random walk model, and time-correlated process model on tracking accuracy was investigated in [74]. The results suggested that the choice of the dynamics model should account for the silence gap in speech sources and continue tracking the sound sources within the silence period as most speakers will not make abrupt changes in their direction and velocity. The tracking system can be made to be more robust against the effects of noise and reverberation if the assumption that the speakers retain their original trajectory prior to the silence period is incorporated in the dynamics model.

The dynamics model used to propagate a single speaker in the single state space is the coordinate uncoupled Langevin model [64] [75] [74]. Assuming the position vector of sound source in a single dimension is defined as $\mathfrak{p} = [\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}},k}\ \mathfrak{p}_{\mathrm{y}_{\mathrm{coord}},k}]^T$, at time, k, the following equation shows the Langevin model for the single dimension $\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}},k}$:

$$\dot{\mathfrak{p}}_{\mathrm{x}_{\mathrm{coord}},k} = a_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}} \dot{\mathfrak{p}}_{\mathrm{x}_{\mathrm{coord}},k-1} + b_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}} u_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}}, \tag{5.4}$$

$$\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}},k} = \mathfrak{p}_{\mathrm{x}_{\mathrm{coord}},k-1} + T_u \dot{\mathfrak{p}}_{\mathrm{x}_{\mathrm{coord}},k}, \tag{5.5}$$

whereby $a_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}}$ and $b_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}}$ are defined as:

$$a_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}} = e^{-B_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}} T_u}, \tag{5.6}$$

$$b_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}} = \bar{v}\sqrt{1 - {a_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}}}^2}, \tag{5.7}$$

with $\bar{v}$ being the steady state velocity and $B_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}}$ being the steady rate constant. T_u is the time interval between each measurement update. $u_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}}$ is a Gaussian variable with zero mean and unit variance,

$$u_{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}}}} \sim \mathcal{N}(0, 1). \tag{5.8}$$

Equations 5.5 and 5.4 apply to each spatial dimension being tracked so the single unlabelled state, x_k, contains the location and velocity information at time k:

$$x_k = \{\mathfrak{p}_{\mathrm{x}_{\mathrm{coord}},k}, \mathfrak{p}_{\mathrm{y}_{\mathrm{coord}},k}, \dot{\mathfrak{p}}_{\mathrm{x}_{\mathrm{coord}},k}, \dot{\mathfrak{p}}_{\mathrm{y}_{\mathrm{coord}},k}\}^T \tag{5.9}$$

Given the transition matrix F is defined as,

$$F = \begin{bmatrix} 1 & 0 & aT_u & 0 \\ 0 & 1 & 0 & aT_u \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix}, \tag{5.10}$$

and the variance Q is defined as,

$$Q = \begin{bmatrix} b^2{T_u}^2 & 0 & 0 & 0 \\ 0 & b^2{T_u}^2 & 0 & 0 \\ 0 & 0 & b^2 & 0 \\ 0 & 0 & 0 & b^2 \end{bmatrix}. \tag{5.11}$$

The state transition equations at a given time k, can be expressed as a matrix equation of the following form:

$$x_k = Fx_{k-1} + u_k, \tag{5.12}$$

with the dynamical noise u_k, defined as:

$$u_k \sim \mathcal{N}(\mathbf{0}, Q), \tag{5.13}$$

where $\mathbf{0}$ is $\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^\intercal$. Given the equations 5.12 and 5.13 the single-target state transition kernel is

$$f(x_k|x_{k-1}) = \mathcal{N}(x_k, Fx_{k-1}, Q). \tag{5.14}$$

5.3.2 Observation Models

In a Bayesian tracking framework, the likelihood function measures the probability of receiving a measurement, z, given the state, x. In the original work using PF to track a sound source [64], the likelihood function for TDoA estimation based localisation technique is expressed as

$$g(z_k|x_k) = q_0 + \sum_{\varpi=1}^{N_\varpi} q_\varpi \mathcal{N}\left(\Delta; \hat{\Delta}^\varpi, \sigma^2\right) \tag{5.15}$$

where $\mathcal{N}$ is the Gaussian distribution, N_ϖ is the number of sources detected, Δ is the TDoA of the particle and $\hat{\Delta}$ is the mean TDoA observed by a sensor pair.

The mean TDoA, $\hat{\Delta}$, observed by a microphone pair is obtained from the relative delay, $\tilde{\delta_{n,k}}$. The relationship between the relative delay and the TDoA is described in equation (5.16).

$$\Delta_{n,k} = \frac{\tilde{\delta_{n,k}}}{f_\mathrm{s}} \tag{5.16}$$

whereby $\Delta_{n,k}$ refers to the true TDoA of a nth peak at time k, $\tilde{\delta_{n,k}}$ is relative delay and f_s is the sampling frequency. As mentioned in Section 2.5.1, equation (2.26) is used [76] to establish a relationship between the observed measurement and the state estimate

$$\hat{\Delta}_k = \frac{|\hat{\mathfrak{p}}_{s,k} - \mathfrak{p}_{\mathrm{mic},m}| - |\hat{\mathfrak{p}}_{\mathbf{s},\mathbf{k}} - \mathfrak{p}_{\mathrm{mic},\mathbf{n}}|}{c} \tag{5.17}$$

with $\hat{\Delta}_k$ referring to the estimated TDoA between the microphone pair at time k, $\hat{\mathfrak{p}}_{s,k}$ being the estimated location vector of the sound source which contains the XY co-ordinates, $\mathfrak{p}_{\mathrm{x_{coord}},k}$ and $\mathfrak{p}_{\mathrm{y_{coord}},k}$, $\mathfrak{p}_{\mathrm{mic},m}$ and $\mathfrak{p}_{\mathrm{mic},n}$ referring to the microphone pair and c is the speed of sound.

In order to account for the presence of the sound source, q_0 is the prior probability that none of the TDoA is due to the true source while q_ϖ is the probability that the ϖth TDoA resulted from the true source. The values for both q_0 and q_ϖ range from 0 to 1. In cases when there is no prior information on the likely source locations, q_ϖ can be expressed as:

$$q_\varpi = \frac{1 - q_0}{N_\varpi}, \qquad \varpi \in \{1, \cdots, N_\varpi\} \tag{5.18}$$

In CBMeMBer, a Gaussian likelihood model is used as the likelihood function because of the acoustic feature extraction method and the dynamics model chosen. It is assumed that the sensor measurements and the dynamics of the sound source are

corrupted by a Gaussian noise. However in the proposed solution, both the variables q_0 and q_ϖ which accounts for a target's birth and survivability in equation 5.15 are dropped from the likelihood function. This is because the CBMeMBer already accounts for the targets' birth and death through the use of a Poisson RFS. In the observable space, the likelihood of the single target state given the measurement is a Gaussian with the mean being the TDoA, $\Delta_{n,k}$, obtained from relative delay of the TF-weighted histogram peaks

$$g_k(z|x) = \mathcal{N}(\Delta_k, \hat{\Delta}_k, \sigma_\Delta^2) \tag{5.19}$$

with Δ_k being the observed instantaneous TDoA at time k used as the observation, z, while $\hat{\Delta}_k$ is the calculated TDoA from the state, x, and σ_Δ being the standard deviation from the observed TDoA. The set of states, X, used in the CBMeMBer contains the position and velocity information while the observed measurement, z, is in terms TDoA, Δ_k. The acoustic features extracted by the BSS framework are the relative attenuation, $\tilde{a_{n,k}}$, and the relative delay, $\tilde{\delta_{n,k}}$. Only the relative delay is used as a measurement in the CBMeMBer. The relative delay of a nth peak, $\tilde{\delta_n}$,is transformed back into real TDoA before being used in the CBMeMBer filter as observations for a particular frame.

5.3.3 Data fusion in CBMeMBer

With only a single pair of microphone, only DoA of the speech sources can be estimated from the measurements. In order to estimate the Cartesian Coordinates of the speech sources, multiple readings are required. With the alteration made to the BSS via signal sparsity algorithm, acoustic features from multiple pairs of microphones can be extracted. The CBMeMBer algorithm has to be adapted to fuse the readings from multiple sensor pairs. For the CBMeMBer filter, the iterated-corrector approximation technique is used to fuse data from multiple sensors. Although the iterated-corrector approximation method of approximating the multi-target Bayes posterior density is not ideal as the order in which the sensor pairs are processed will affect the result, it is stated in [110, p.595] that the use of iterated-corrector approximation method does not result in noticeable degradation of performances during simulations and it is adequate for practical applications [120]. The multi-target Bayes posterior density is approximated by recursively updating the predicted likelihood with the measurements from each sensor pairs. The multi-target posterior density is proportional to the product of

the likelihood as expressed in equation (5.20)

$$\pi_k(X_k|Z^1_{1:k},\cdots,Z^M_{1:k}) \propto \pi_{k|k-1}(X_k|Z^1_{1:k-1},\cdots,Z^M_{1:k-1}) \prod_{m=1}^{M} g_{k,m}(Z^m_k|X_k) \quad (5.20)$$

with $\pi_k(X_k|Z^1_{1:k},\cdots,Z^M_{1:k})$ referring to the multi-target Bayes posterior density, $g_{k,m}(Z^m_k|X_k)$ referring to the mth sensor pair's observation and
$\pi_{k|k-1}(X_k|Z^1_{1:k-1},\cdots,Z^M_{1:k-1})$ referring to the predicted multi-target density.

5.4 Integration of the BSS via Signal Sparsity framework with CBMeMBer

The combination of BSS via Signal Sparsity and CBMeMBer allows multiple moving speech sources to be localised and tracked but the estimated information is still not enough for speech source separation. The CBMeMBer filter only estimates the tracks of the speech sources but it does not associate the tracks with the speech sources. A track management extension is required to assign labels to the tracks estimated by the CBMeMBer as it does not incorporate labels as part of the parameters propagated. As the source separation stage is part of the BSS via signal sparsity technique, the proposed solution requires TF masks to perform speech source separation. The TF masks are constructed based on the acoustic features estimated from the speech sources' tracks and their associated labels. By applying the TF masks onto the speech mixture, the individual speech sources can be estimated in the TF domain. The separated speech sources can be reconstructed in the time domain by applying an Inverse Short Time Fourier Transform (ISTFT) on these separated speech sources. The process of source separation and reconstruction is illustrated in figure 5.4.

5.4.1 Track Management

Target state estimation is performed by the CBMeMBer technique without the need for data association in the filtering process [113]. By not incorporating labels as part of the parameters propagated in the multi-Bernoulli approximation, the computational complexity to run the CBMeMBer algorithm remains low. It is computational expensive to estimate labels within the RFS framework so the CBMeMBer does not inherently incorporate data association in its algorithm. However, the identity information provided

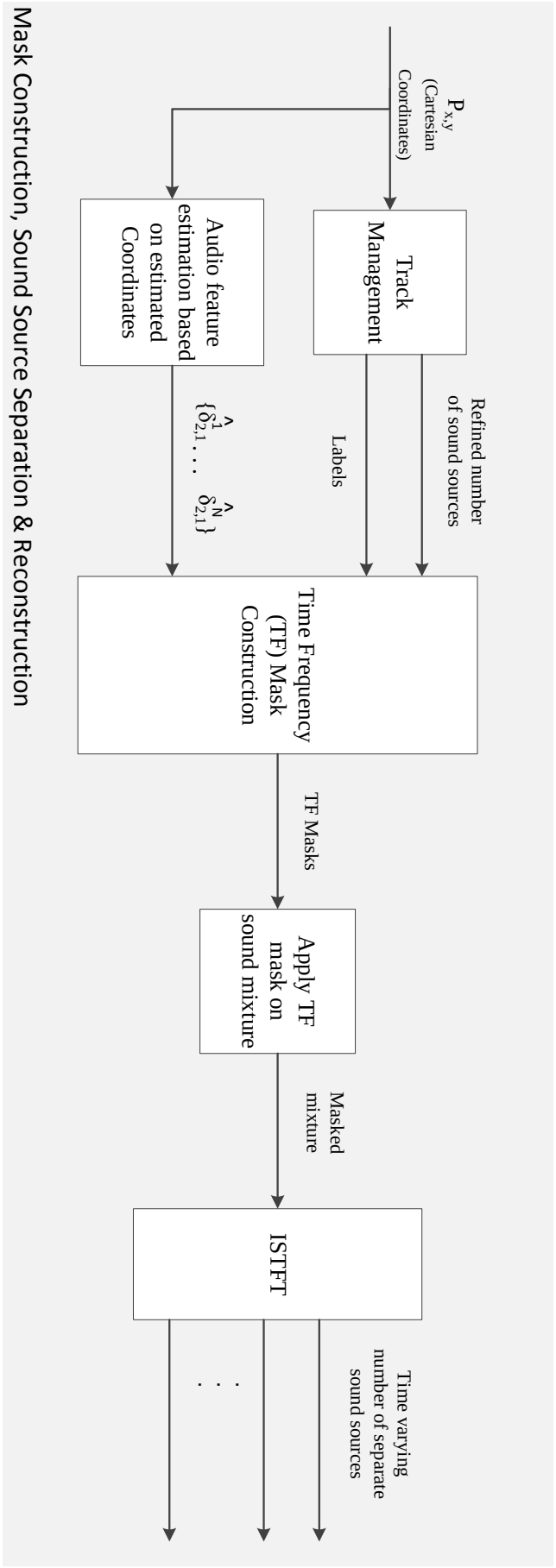


Figure 5.4: Block diagram of the Mask Construction and Source Separation Process

by data association is required to construct the TF masks used for speech source separation. In order to assign identity information to the hypothesized tracks corresponding to the speech sources while retaining the tractability of the CBMeMBer filter, a separate track management algorithm is required. A separate track management extension will allow the system more flexibility in using the past history of the state estimates to assign identity information to the hypothesized tracks representing the speech sources. The track management system used in this research is similar to the ones used in [114] and [113] with modifications made to suit acoustic signal analysis. The focus of the track management extension used in this research is to use the state estimates from the CBMeMBer as input and outputs the labels and the refined number of speech sources based on the number of labels assigned. This process is illustrated in figure 5.5.

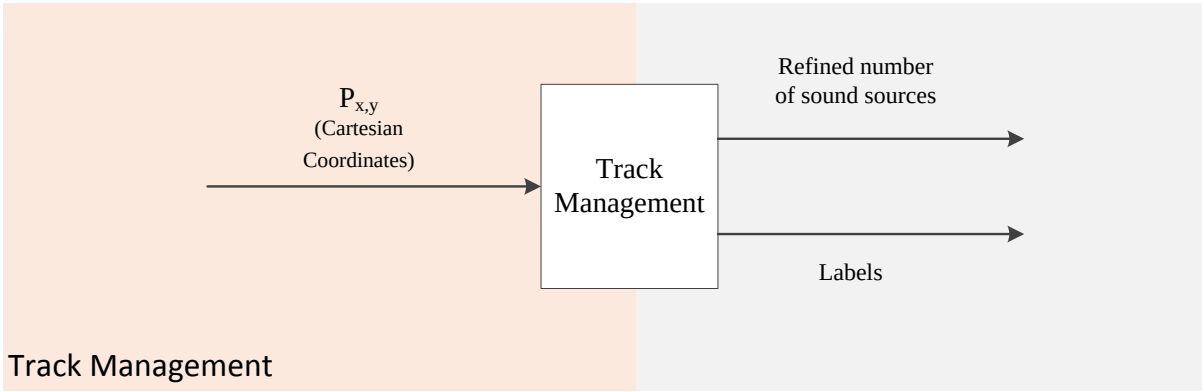


Figure 5.5: Block diagram of the Track Management Process

For the purpose of assigning identity information to the tracked speech sources, there are multiple factors such as silence period, speaker interaction and room reverberation that needs to be considered. In order to account for these factors in speech source tracking, certain constraints are placed on the track management system. Due to the nature of speech source, there might be silence periods which provide no information that allows the it to be tracked. As a result of this, the track management system has to be able to retain the track and identity information for the particular speech source instead of terminating it. By applying the idea of the track management used in [121], the labels of several time steps are taken into consideration when identities are assigned to the speech sources. As discussed in Section 4.3, the original labelling extension proposed in [13] only looks at the past time step for data association and it performs poorly when two targets are merged in one frame and separated in the next.

When such a scenario occurs, one of the sound sources will be assigned a new label instead of its original label prior to the merging. Sound sources lie on a hyperbolic surface [76] so there are occasions when two sound sources lie on the same hyperbolic surface and produce the same TDoA. In such a scenario, one of the speech sources will be occluded by the other speech source and the two targets are ”merged” in terms of measurement despite not physically merged. These “merging” problems can be solved through the use of a memory which retains the past information of these states [113].

The track management based on [121] has four basic stages. The first stage is to associate the estimated data at current time step with the data from the previous time step. Non-associated data in current time step is then considered for data association with previously missed data. If the current data has no association with the data from previous time steps, a new identity is assigned to the current data. Memory retention will only be available for a set number of time steps. If a data remains dissociated for a certain time period, it will be removed from the data association memory. In the cases of moving sound sources, the location of the sound sources will change based on the trajectories of the individual sound sources. The true estimates will follow a trajectory whereas the false estimates which result from noise or reverberation will be spurious. Hence, the true tracks will be considered for data association while the spurious tracks that result from noise or reverberation will be removed from track association.

A modification made to the track management system is the addition of a track merging algorithm in the CBMeMBer filter. Tracks within a short distance of each other are merged into a single track as such scenarios are highly unlikely in real life. There will always be a personal space between speakers in most social settings so the boundary of this personal space is used as the threshold for track merging [122]. As the track management accounts for missed data, the merged track will not be lost as long as the tracks do not overlap for a period of time exceeding the memory retention period of the track management algorithm. The details of the track management algorithm is given in pseudocode 4.

Algorithm 4 Pseudocode for Multiple Speech Sources Track Management

Data Initialization

Acquire the estimated speech source locations from the CBMeMBer filter starting from the first time frame
Set the track merging threshold, association distance threshold and maximum memory retention period
Set all association bits for the estimated sound source locations to 1
Set all missed bits for the estimated sound source locations to 0
Assign a new label for each estimated sound source location

Track Merging

Compare the distance for all the estimated sound source location at each time frame
while Distance between two estimated locations $<$ merging threshold **do**
    Combine the particle clouds of those two estimates
    Reweight the particles
    Combine the probability of existence of the two particle clouds
end while
Estimate new sound source locations based on the combined particle clouds

Iterative Data Association

while time frame $\neq$ last time frame **do**
    Compare estimates from current time frame with the directly previous time frame
    if Distance between current estimate and previous estimate is within the association distance threshold **then**
        Assign the previous label to current estimate and set the corresponding association bit to 1
    end if
    Compare estimates from current time frame with previously missed data
    if Distance between current and the previously missed estimate is within the association distance threshold $*$ missed time steps **then**
        Assign the previously missed estimate's label to current estimate and set the corresponding association bit to 1
    end if
    if Current estimate has no association to previous data or previously missed data **then**
        Assign a new label to the current estimate and set the corresponding association bit to 1
    end if
    if Previously missed data $>$ maximum memory retention period **then**
        Increase its missed estimate counter by 1
    else if Previously missed data $>$ maximum memory retention period **then**
        Remove it from memory
    end if
end while

5.4.2 Acoustic Features Estimation from Source Location

In conventional BSS via signal sparsity techniques, the TF mask construction relied on the true acoustic features of each sound sources estimated from the clustering techniques. The acoustic features belonging to each sound source were estimated from the peak of the power weighted histogram or the centroids of the clusters. This technique was valid as the speech energy for the non-stationary and non-moving sound sources was concentrated in the regions of the power-weighted histogram peaks [4]. In the case of non-stationary moving sound sources, the peaks do not necessarily represent the sound sources as the histograms tend to be merged as the acoustic features change relative to the position of the speech sources. The proposed method of estimating the true acoustic features for moving speech sources is to use the track management to provide identity association for the different estimated tracks and extract the true acoustic features of the moving speech sources from the coordinates of these tracks. The acoustic feature estimation process is illustrated in figure 5.6.

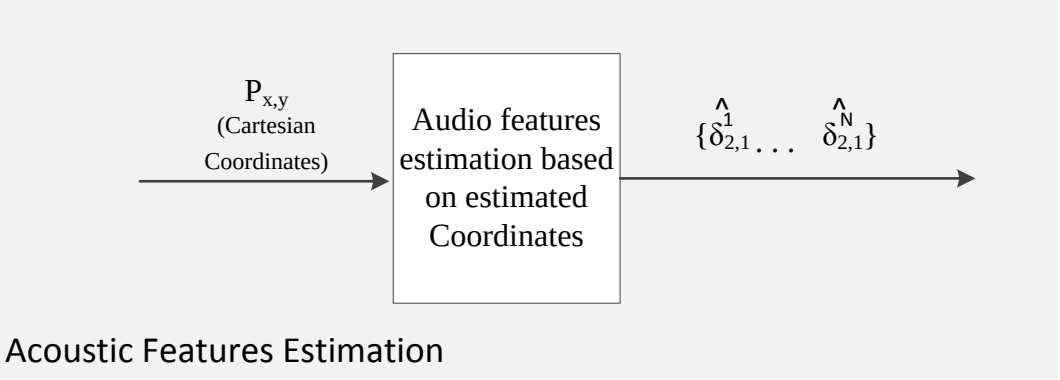


Figure 5.6: Block diagram of the Acoustic Features Estimation Process

In conventional BSS via signal sparsity techniques, the acoustic features required for the mask constructions are attenuation and delay. With a one dimensional tracking in terms of just DoA, only the TDoA of the sound source can be estimated. However, both the relative attenuation and the relative delay can be estimated if the positions of the sound source and the sensor pairs are known. The relationship between the sound source and the acoustic features is illustrated in figure 5.3. The reason of estimating the acoustic features from the labelled tracks instead of the raw output from the BSS framework is twofold. First, such a technique allows the acoustic features of a moving speech source which is changing relative to the position of the speech source to

be estimated. Second, the spatio-temporal relationship established for the true sound sources allows the CBMeMBer to mitigate the effects of noise and reverberation on the accuracy of the acoustic features estimates.

In [76], it is shown that the TDoA can be estimated from the source locations. With reference to figure figure 5.3, the TDoA of a speech source can be estimated from the position of the speech source as shown in equation (5.21).

$$\Delta_n = \frac{d}{c} = \frac{r_2 - r_1}{c} \tag{5.21}$$

with Δ_n referring to the TDoA, d referring to the distance between the microphone pair, r_2 and r_1 referring to the distance between the sound source with the second and the first microphones respectively and c refers to the speed of sound which is assumed to be $343m/s$. Equation (5.21) is just equation (2.26) expressed in Cartesian Coordinate form. By rearranging equation (5.16), the TDoA obtained can be converted back to the relative delay to be used in the construction of the TF mask through the following formula

$$\hat{\delta_n} = \Delta_n \times f_{\mathrm{s}} \tag{5.22}$$

whereby $\hat{\delta_n}$ and Δ_n refers to the estimated delay and the TDoA of the nth labelled speech source respectively, while f_{s} is the sampling frequency. The TF mask required for source separation is constructed using the identity information from the track management system and the delta estimated from the location of the speech sources.

The proposed solution does not use the estimated attenuation in the construction of the TF masks for the reasons discussed in Section 5.2.2. Although this research does not use attenuation for the construction of the TF mask, the method of estimating the attenuation from the location of the speech source is discussed as the attenuation estimate might be used for other acoustic source problems. The relative attenuation of the sound source is a ratio of the distance between the sound source and the two microphones. The relative attenuation for the nth labelled speech source can be calculated using the equation 5.23.

$$\hat{\alpha_n} = \frac{r_2}{r_1} \tag{5.23}$$

whereby $\hat{\alpha_n}$ refers to the estimated attenuation of a nth labelled sound source. As the proposed method directly estimates the attenuation from the labelled tracks, there is no

need for a conversion of the symmetric attenuation back into the relative attenuation as discussed in Section 3.2.3.

5.4.3 Mask Construction and Separated Sound Source Reconstruction

An integral part of the BSS via signal sparsity technique is the use of TF masks to achieve sound source separation by masking sound sources aside from the sound source of interest. The TF masks used to separate the speech sources require the identity information of the speech sources. The original TF masks which was used to separate non-moving sound sources, relies on the acoustic features estimated from the peak of the power-weighted histogram or centroids of acoustic feature clusters for identity information. The number of masks created was based on the observed number of peaks or clusters. When the speech sources are moving, the method of relying on peaks or clusters for identity is no longer possible as the acoustic features changes according to the position of the speech sources. As a result of this, the track management extension was used to assign identity information to the speech sources tracked by the CBMeMBer filter. For the construction of TF masks for moving speech sources, the number of unique labels indicate the number of masks to be created and these labels provide an identity association for moving speech sources. The focus of section 5.4.3 is to utilise the label and target state information to construct the TF masks required for source separation. This process is illustrated in figure 5.7.

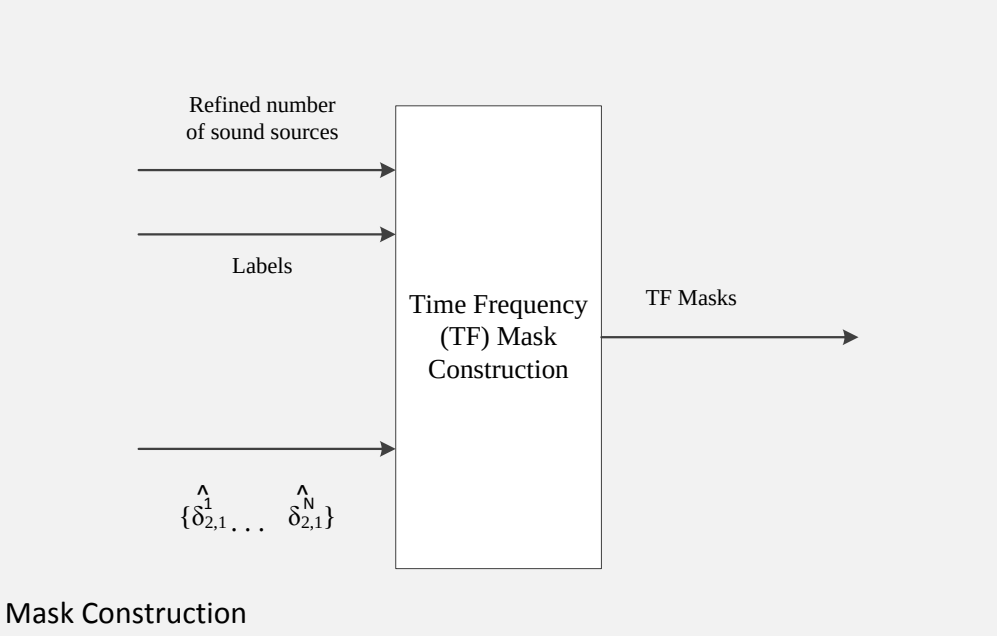


Figure 5.7: Block diagram of the Mask Construction Process

Sound sources reconstructed from binary masks using a hard masking technique have an unnatural "musical noise" in them [104]. A soft masking technique is capable of mitigating the "musical noise" effect on the sound source mixtures. Both [104] and [103] use a fuzzy c means clustering technique to separate the sound clusters and the membership function which determines the likelihood of the acoustic features belonging to a particular cluster is directly used as the mask. Taking inspiration from [104] and [103], a soft mask is proposed for this research. The proposed mask is constructed using a normal Gaussian distribution with the estimated acoustic features as the mean and the measurement variance as the variance of the distribution. The construction of TF mask using a normal Gaussian distribution is motivated by the modeling assumptions used for the dynamics and likelihood model. The dynamics model assumes the actual trajectory of the sound source is corrupted by a Gaussian noise while the likelihood model assumes the measurements observed are a normal Gaussian distribution. The proposed TF mask is expressed below in equation (5.24)

$$\tilde{M}_n(\tau, \omega) := \mathcal{N}(\delta, \hat{\delta_n}, \sigma_\delta) \tag{5.24}$$

with δ being the instantaneous delay and attenuation of each TF bin in the speech source mixture while $\hat{\delta_n}$ is the estimated delay of a nth sound source as indicated by the label of the speech source's trajectory.

After the TF masks are constructed, the masks are applied to the sound mixture in order to separate the speech sources. The process of speech separation and reconstruction is illustrated in figure 5.8.

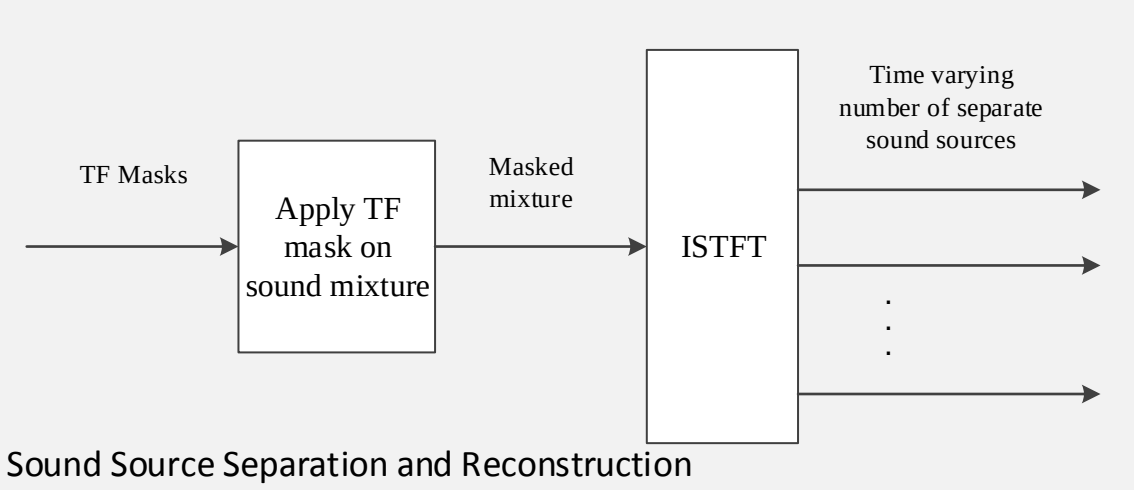


Figure 5.8: Block diagram of the Source Separation and Reconstruction Process

The TF masks are applied to the sound source mixture according to the following equation

$$\hat{S}_n(\tau, \omega) = \tilde{M}_n(\tau, \omega) Y_o(\tau, \omega) \tag{5.25}$$

with $\hat{S}_n(\tau, \omega)$ referring to the estimated nth labelled sound source and Y_o refers to the corresponding signal mixture received by one of the microphone pair used for the acoustic features extraction. Although multiple pairs of microphones were used in the localisation and tracking of the speech sources, only a pair is required to extract the acoustic information and to reconstruct the signal. It is important to note that the signal mixture which is to be masked has to correspond to the microphone pair used for the acoustic features extraction. This is due to the fact that each microphone pairs will have a different attenuation and delay relative to the speech source's position. If the acoustic features used to construct the mask and the sound mixture do not match, the estimated sound source will be incorrect.

In this research, the microphone pairs are placed across the room in order to achieve spatial diversity in the TDoA measurements received. The spatial diversity allows the Cartesian Coordinates of the speech sources within the boundaries of the room to be estimated rather than just the DoA. It is understood that in the real environment, the distance between the speech sources to the microphone will affect the power of the

signal received by the microphone pair. The source mixture from microphone pair selected for masking is best chosen based on the proximity of the speech source to the particular sensor and the power of the signal received. However, the problem of microphone pair selection for enhanced quality of separated speech source is a challenging problem on its own and it is beyond the scope of this thesis. This thesis is a proof of concept that a time varying number of speech sources can be tracked and separated using the integration of BSS via signal sparsity and CBMeMBer techniques. Hence, only the speech source mixture from the reference microphone is selected to be masked.

The masked sound sources are in the TF domain. The final step in the source separation for moving speech sources process is to use an Inverse Short Time Fourier Transform (ISTFT) to reconstruct the separated moving speech sources in the time domain. The whole process of recursively localising, tracking and separating moving speech sources online is outline in algorithm 5.

5.4.4 Summary

In summary, Chapter 5 outlines the main contribution of this thesis which is the adaptation of the BSS via signal sparsity technique for online speech source separation and integrating it with the CBMeMBer target tracking technique in order to solve the problem of source localisation, tracking and separation for multiple moving speech sources.

The proposed adaptations to overcome the limitations of the original BSS via signal sparsity algorithm in the separation of moving speech sources was discussed in Section 5.2. The main adaptation made to the BSS algorithm for integration with CBMeMBer was to process the source mixture on a frame by frame basis and extract acoustic features for each time frame. The other adaptations made was the use of the audio signal's cross power spectral density over its auto power spectral density to calculate the TF ratio and the use of only the relative delay for sound source localisation as measurement in the CBMeMBer filter. The final adaptation made for the BSS algorithm is the use of multiple pairs of microphones in order to achieve source localisation in two dimensions (Cartesian Coordinate).

In order to adapt the CBMeMBer filter for speech source tracking in acoustic scenarios, suitable dynamic and likelihood models are required. The single target dynamics

Algorithm 5 Pseudocode for Source Localisation, Tracking and Separation of Moving Speech Sources

Acoustic Feature Extraction

1: Extract the acoustic feature - delay, $\tilde{\delta_n}$, using the BSS via signal sparsity technique.
2: Transform the relative delay into TDoA using the equation $\Delta_n = \frac{\tilde{\delta_n}}{f_s}$.

State Estimation

3: Predict the multi-target prior density, $\pi_{k|k-1}$, at time k based on the multi-target posterior density at time $k-1$. The multi-target prior density is a union between the surviving multi-Bernoulli parameter set with the target birth: $\pi_{k|k-1} = \{(r^{(i)}_{P,k|k-1}, p^{(i)}_{P,k|k-1})\}_{i=1}^{M_{k-1}} \cup \{(r^{(i)}_{\Gamma,k}, p^{(i)}_{\Gamma,k})\}_{i=1}^{M_{\Gamma,k}}$
4: Update the predicted prior density with the received measurement from BSS via signal sparsity algorithm, t_j. Propagate the CBMeMBer's *parameterised approximation of the multi-target Bayes posterior density*, $\pi_k \approx \{(r^{(i)}_{L,k}, p^{(i)}_{L,k})\}_{i=1}^{M_{k|k-1}} \cup \{(r^{*}_{U,k}(z), p^{*}_{U,k}(\cdot; z))\}_{z \in Z_k}$
5: Estimate the number and the state of the hypothesized tracks

Track Management

6: Based on the estimated state trajectory, perform identity association using the track management extension. Apply unique labels to each of the estimated tracks
7: Estimate the acoustic features for all the labelled tracks. The positions of the speech source and the microphone pair are assumed to be known. The input signal from the microphone pair chosen for the acoustic features extraction must be used for the TF masking.

Mask Construction and Speech Source Separation

8: Construct a TF mask, $\tilde{M}_n$, based on the labels and the acoustic features. A unique mask is created for each label. The proposed soft mask is a Gaussian distribution of the estimated delay - $\tilde{M}_n(\tau, \omega) := \mathcal{N}(\delta, \hat{\delta}_n, \sigma_\delta)$.
9: Source separation is achieved by applying the TF mask on the sound mixture received from one of the microphones used for acoustic features estimation - $\hat{S}_n(\tau, \omega) = \tilde{M}_n(\tau, \omega) Y_o(\tau, \omega)$.
10: Reconstruct the separated speech sources in time domain by transforming the masked mixtures using ISTFT.

model used is the Langevin model as human speakers do not always move in a linear fashion. The likelihood model is similar to the one used in [64]. Instead of incorporating target survival as part of the likelihood, the CBMeMBer has already modeled the target's birth and death model as a Poisson RFS and recursively propagates it. Hence, the variables which account for target survivability in [64] were excluded from the likelihood model in the proposed solution.

The integration of the BSS via signal sparsity technique and CBMeMBer filter alone is not enough to achieve sound source separation for moving speech sources. A track management system is required to provide data association data or identity for the estimated speech source trajectories. The state information and the identity of the

estimated tracks are used to construct the TF masks used for source separation. The state information of the speech sources is used to estimate the acoustic features while the identity information is used to relate these acoustic features across time frames. By applying the unique TF masks on the speech source mixture, speech source separation is achieved.

Chapter 6

Experiments and Discussion

There are no bad restaurants or bad experiments in this world.

– Tendou Souji

6.1 Overview

In Chapter 2, the problem of localising, tracking and separating multiple moving speech sources was outlined. An online iterative solution that utilises the signal sparsity property of sound source mixtures and the target tracking technique of CBMeMBer was proposed to solve the source separation problem in a "conference room scenario". The background of the BSS via signal sparsity was discussed in Chapter 3 while the background of CBMeMBer was discussed Chapter 4. Chapter 5 details the contribution of this thesis which is the proposal of an online recursive solution the source separation problem in the "conference room scenario". The proposed solution adapts the BSS via signal sparsity approach to extract the acoustic features for moving speech sources and to separate the speech sources. The CBMeMBer filter has also been adapted to estimate the number of speech sources and to track these speech sources based on the extracted acoustic features. In order to assign identities for the tracked speech sources, a track management extension is applied after the CBMeMBer process. The identity information is used to associate the acoustic features with their respective speech sources and to construct the TF mask. Moving speech sources separation is achieved by applying the unique TF masks onto the speech source mixtures. With first five chapters explaining the details of the proposed solution, the focus of Chapter 6 is

the evaluation of the proposed solution.

6.2 Evaluation Criteria

The main contribution of this thesis is to propose a solution which is capable in tracking and separating multiple moving speech sources. Hence, the proposed solution needs to be evaluated against several criteria based on the functions it has to perform. In terms of localisation and tracking, the Optimatal SubPattern Assignment (OSPA) metric [123] was used to evaluate it. The *BSS EVAL Toolkit* developed by Vincent in [124] was used to evaluate the performance of speech source separation.

6.2.1 Evaluation Criteria for State Estimation

The evaluation the tracking results of the proposed solution was performed using the OSPA metric [123]. As discussed in Section 2.5.2 and Chapter 4, the CBMeMBer filter used to track the speech sources is an RFS filter which considers the tracked targets as a multi-target state. The number of the speech sources states as well as the states of these tracked speech sources evolve with time so the multi-target miss distance has to be able to account for the differences between the reference multi-target state and the estimated multi-target state [123]. Hence the OSPA metric, which is a metric on the space of finite sets and is capable of capturing both the state and cardinality errors, is chosen as the metric to evaluate the tracking performance of the proposed solution. Apart from that, the OSPA metric is easily computed and it can be intuitively interpreted [123].

The definition of the OSPA metric $\bar{d}_p^{(c)}$ is given by denoting $\bar{d}^{(c)} := \min(c, |x - y|)$ for $x, y \in \mathcal{X}$ with the cut off at $c > 0$ and $\prod_k$ the set of permutations on $\{1, 2, \ldots, k\}$ for any positive integer k. Given $p \geq 1, c > 0$, and both the sets $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ are on the multi-target finite space, $\mathcal{F}(\mathcal{X})$, the OSPA metric

can be calculated as [73, p.94]

$$\bar{d}_p^{(c)}(X,Y) := \begin{cases} 0, & m = n = 0, \\ \left(\frac{1}{n}\left(\min\limits_{\pi\in\prod_n} \sum_{i=1}^{m} d^{(c)}(x_i, y_{\pi(i)})^p + c^p(n-m)\right)\right)^{\frac{1}{p}}, & m \leq n, \\ \bar{d}_p^{(c)}(Y,X), & m > n. \end{cases} \tag{6.1}$$

The OSPA metric is defined between two point pattern and it is equal to c if one of the set is empty while the other is nonempty.

When used to evaluate the performance of multi-target tracking, the OSPA metric is interpreted as a pth order per target error and it is made up of two components - the pth order per target localisation error and the pth order per target cardinality error. However, these error components are not metrics in the space of finite subsets. The purpose of the order parameter p is to adjust the sensitivity of the metric to outliers in the data while the purpose of cut off parameter c is to determine the relative weighting of how the metric penalises the cardinality error against the localisation error. The pth order per target localisation error can be expressed as

$$\bar{e}_{p,\text{loc}}^{c} = \left(\frac{1}{n}\min_{\pi\in\prod_n} \sum_{i=1}^{m} d^{(c)}(x_i, y_{\pi(i)})^p\right)^{\frac{1}{p}} \tag{6.2}$$

while the pth order per target cardinality error can be expressed as

$$\bar{e}_{p,\text{card}}^{c} = \left(\frac{c^p(n-m)}{n}\right)^{\frac{1}{p}}. \tag{6.3}$$

In most cases, it is sufficient to use the OSPA metric to evaluate the performance of a multi-target tracking algorithm but the $\bar{e}_{p,\text{loc}}^{c}$ and $\bar{e}_{p,\text{card}}^{c}$ may be able to provide additional insight into the tracking performance.

6.2.2 Evaluation Criteria for BSS

The speech separation performance is evaluated using the evaluation criteria developed by Vincent et al in [125] as these criteria can be computed for all separation algorithms and do not require the information of the separation filters or masks. The separated signals are defined as a linear composition of the target signal, the interference signal, the artificial signal and noise as shown in equation 6.4 [124]

$$\hat{s}_n(t) = s_{\text{tar},n}(t) + e_{\text{int}}(t) + e_{\text{art}}(t) + e_{\text{spat}}(t) \tag{6.4}$$

with $\hat{s}_n(t)$ being the estimated n^{th} speech source and $s_{\text{tar},n}(t)$ being a version of the original n^{th} speech source, $s_n(t)$, modified by some allowable distortions. Hence, the original n^{th} sound source can be approximated as $s_n(t) \approx s_{\text{tar},n}(t)$. The interference, artifact and spatial error terms are represented by $e_{\text{int}}(t)$, $e_{\text{art}}(t)$ and $e_{\text{spat}}(t)$ respectively. In [125], the spatial distortion and the interference components are computed by projecting the least squares of estimated source signal onto the corresponding subspaces. The interference, artifact and spatial error terms are expressed as

$$e_{\text{int}}(t) = \mathcal{P}^L_{\text{all}}\hat{s}_n(t) - \mathcal{P}^L_n\hat{s}_n(t) \tag{6.5}$$

$$e_{\text{art}}(t) = \hat{s}_n(t) - \mathcal{P}^L_{\text{all}}\hat{s}_n(t) \tag{6.6}$$

$$e_{\text{spat}}(t) = \mathcal{P}^L_n\hat{s}_n(t) - s_{\text{tar},n}(t) \tag{6.7}$$

whereby $\mathcal{P}^L_n$ is the least squares projector onto the subspace spanned by $s_{\text{tar},on}(t-\tau)$, $1 \leq o \leq M$, $1 \leq \tau \leq L$ while $\mathcal{P}^L_{\text{all}}$ is the least squares projector onto the subspace spanned by $s_{\text{tar},op}(t-\tau)$, $1 \leq o \leq M$, $1 \leq p \leq N$, $1 \leq \tau \leq L$ [125]. L is the filter length.

The energy criteria laid out in [125] evaluates the separated speech sources in terms of Image-to-Source Distortion Ratio (ISR), Source-to-Interference Ratio (SIR) and Source-to-Artifact Ratio (SAR). The relative amount of spatial distortion present in the n^{th} speech source is defined by the ISR. The ISR is calculated as

$$\text{SIR}_n = 10\log_{10}\frac{\sum_{m=1}^{M}\sum_t s_{\text{tar},n}(t)^2}{\sum_{m=1}^{M}\sum_t e_{\text{int}}(t)^2}. \tag{6.8}$$

The SIR which is a measure of the interference present in the n^{th} source signal is expressed as

$$\text{SIR}_n = 10\log_{10}\frac{\sum_{m=1}^{M}\sum_t s_{\text{tar},n}(t)^2}{\sum_{m=1}^{M}\sum_t e_{\text{int}}(t)^2}. \tag{6.9}$$

The relative amount of artifacts present in the n^{th} source signal is defined by the SAR and it is expressed as

$$\text{SAR}_n = 10\log_{10}\frac{\sum_{m=1}^{M}\sum_t (s_{\text{tar},n}(t) + e_{\text{int}}(t) + e_{\text{spat}}(t))^2}{\sum_{m=1}^{M}\sum_t e_{\text{art}}(t)^2}. \tag{6.10}$$

In [125], the total error of the n^{th} source signal is given by the Signal-to-Distortion Ratio (SDR). SDR is a ratio of the source to the total error and it can be expressed as

$$\text{SDR}_n = 10\log_{10}\frac{\sum_{m=1}^{M}\sum_t s_{\text{tar},n}(t)^2}{\sum_{m=1}^{M}\sum_t(e_{\text{int}}(t)+e_{\text{art}}(t)+e_{\text{spat}}(t))^2}. \tag{6.11}$$

Even though there are four criteria laid out to evaluate the performance of speech source separation, the different criterion are used depending on the practical application of the separation algorithm. For this thesis, the SIR and SDR are used to evaluate the separation performance as our interest lies in knowing the amount of interference and error present in the separated speech sources after it has been tracked.

6.3 Preliminary Evaluation

A preliminary evaluation was carried out in [126] to test the viability of using source sparsity to extract acoustic features and tracking the speech sources based on the extracted acoustic features using CBMeMBer. The proposed solution was shown to be viable as it was capable of localising and tracking two speech sources in an ideal scenario with no noise and reverberation. The setup of the room for the priliminary evaluation is shown in figure 6.1. The simulated room is a $10\text{m} \times 10\text{m}$ room with four pairs of microphones which are 0.04m apart, spread out across each of its four walls. The distance between the microphone pairs fulfill the condition of $d < \frac{c}{2f_s}$ as the speed of sound is 343m/s while the frequency of interest is 8kHz. The speech sources used in the scenario are synthetic male and female speech signals sampled at 16kHz. For the settings in the CBMeMBer, the steady state velocity of, $\bar{v}$ and the rate constant, B of a speaker in the Langevin model are respectively set to 1.4m/s and 10Hz. The birth model used is a Gaussian spawning at the birth location of the speakers. The sigma, σ, of the normally distributed likelihood is set at 0.5% of the maximum TDoA between the sensor pair and the clutter rate, κ is set to 6. The result of the preliminary experiment is shown in figure 6.2.

As shown in figure 6.2, the CBMeMBer filter is capable of localising and tracking the speech sources based on the acoustic features extracted by BSS via signal sparsity. The CBMeMBer also correctly estimates the number of speech sources in the scenario. The number of speech sources is represented by the number of tracks. In figure 6.2, the estimates form two distinctive trajectories but there is no identity association to

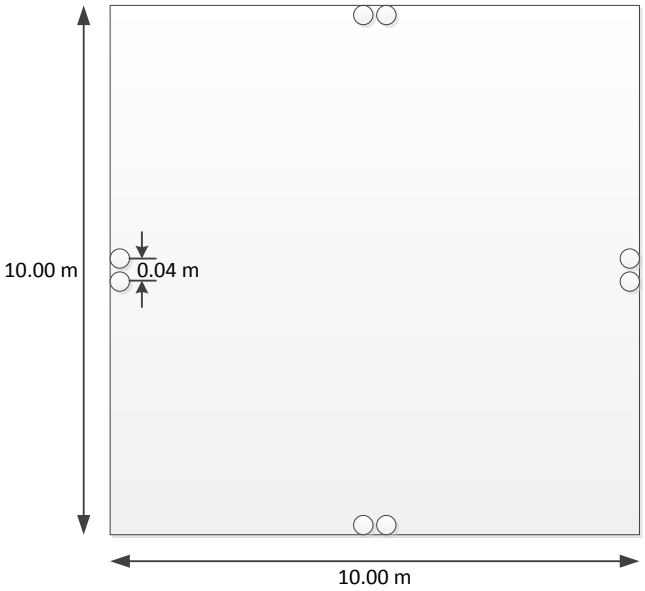


Figure 6.1: Room dimension and setup for preliminary evaluation

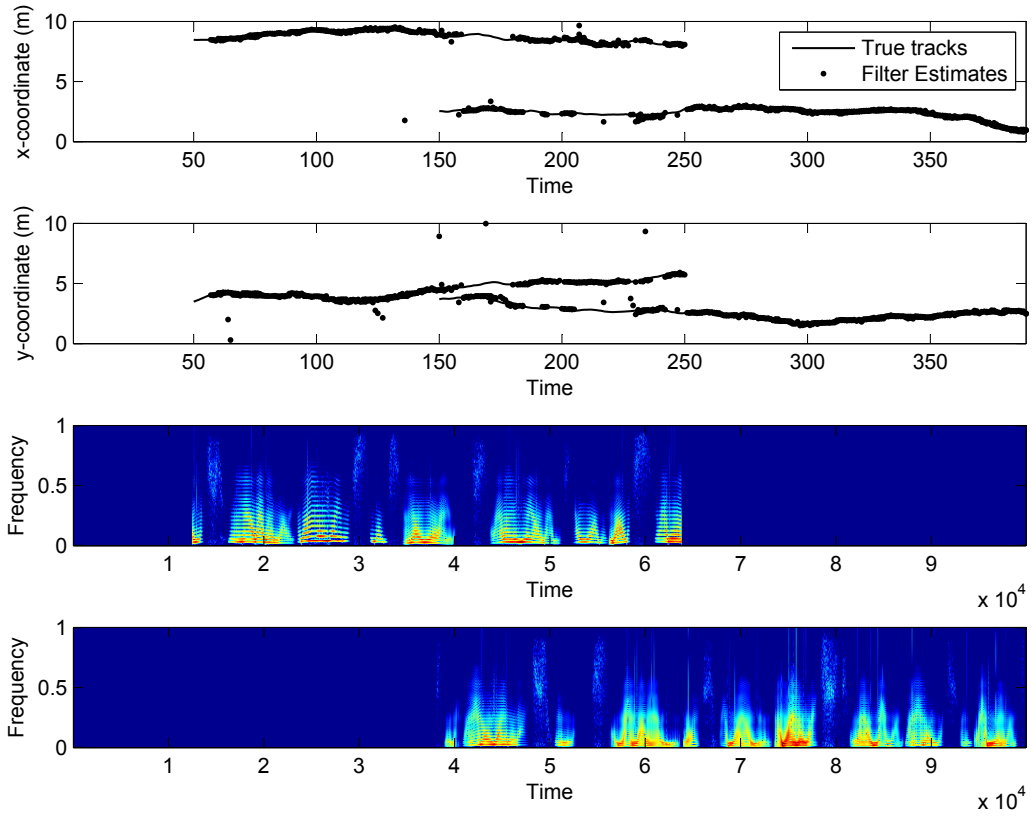


Figure 6.2: Speech source tracking output for two moving sources

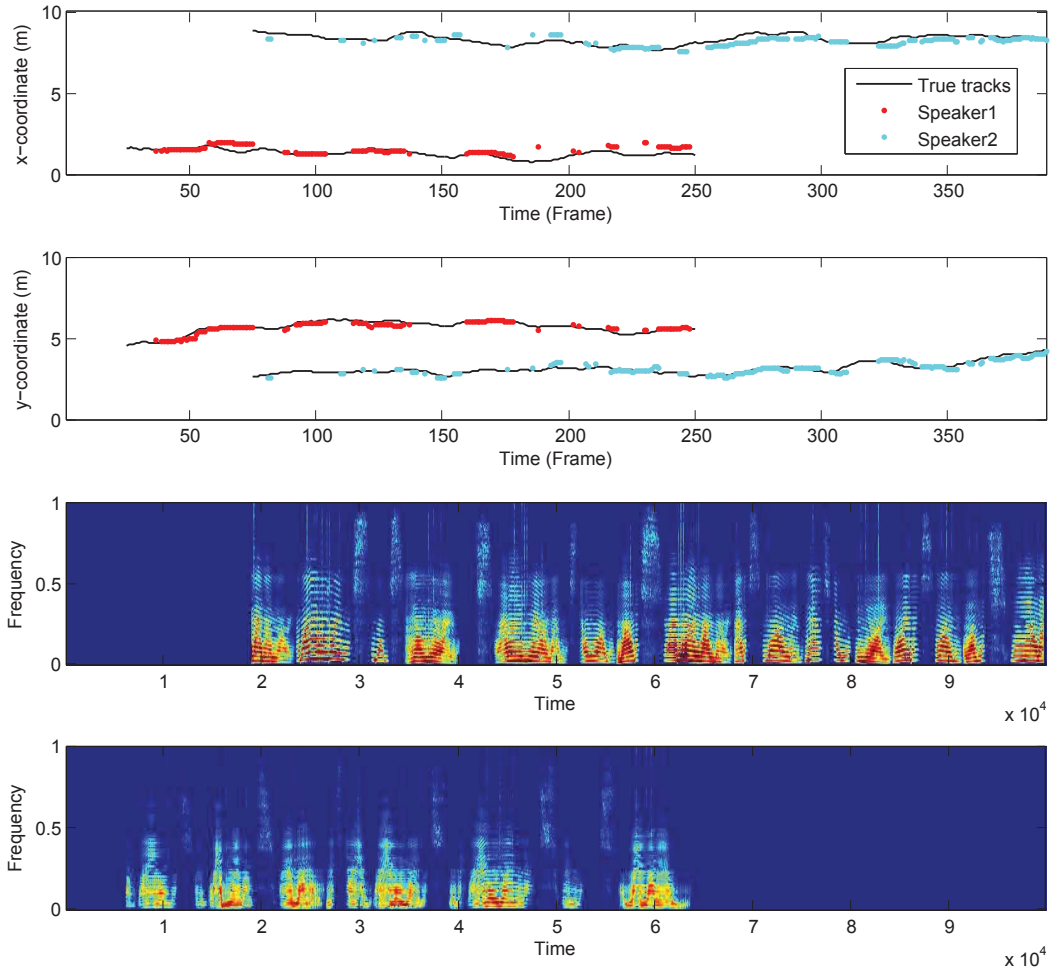


Figure 6.3: Result of source tracking and labelling

separate one trajectory from the other. Hence, the track management algorithm was proposed to extend the capabilities of the CBMeMBer to have explicit labels in [122].

With the addition of the track management algorithm, another evaluation was performed. The simulation environment is similar to the previous room setup. The main difference is the addition of a T60 reverberation time of 0.15s instead of an ideal scenario without any noise or reverberation. The merging threshold used in this simulation is 0.5m so speakers which are within 0.5m of each other are merged into a single track. An example of the simulation results is shown in figure 6.3. As shown in the result, the proposed algorithm is capable of tracking and adding an identity to both the speakers while they are moving. The gaps in the tracks are due to the silent periods in speech. No acoustic features could be extracted during the silent periods so the proposed solution is unable to produce location estimates during these periods. As the track management algorithm is able to account for the silent periods and retain the

track information, the tracks continue to be propagated after these silent periods. As the tracks do not cross path with each other in the scenario shown, it is mainly used to clean up the tracks by eliminating most of the false estimates resulting from room reverberation.

6.4 Experimental Setup

In this thesis, a room simulation with moving speech sources was performed to evaluate the tracking and separation performance of the proposed algorithm in various room environments. The preliminary simulations in terms of tracking were promising so further numerical evaluations were carried out to test the algorithm's robustness in terms of tracking as well as source separation. The effects of different reverberation times, Signal-to-Noise Ratio (SNR), room sizes and mask types were tested out using the proposed algorithm and evaluated using the criteria laid out in Section 6.2.1 and 6.2.2. The room simulations were performed using the code developed by Lehmann [127]. As shown in figure 6.4, the dimension of the room used in this simulation is $8.1\mathrm{m} \times 3.8\mathrm{m}$.

Unlike the previous simulations where four pairs of microphones were laid out in a symmetrical order across the room as shown in figure 6.1, this simulation uses only three pairs of microphones which are spread out across the room in an asymmetrical order. The reason for having an asymmetrical order of microphones is to reduce the number of blind spots for the microphones in order to improve the tracking performance. Symmetrical pairs of microphones which are directly opposite of each other will share the same blind spots. The distance between the microphone pairs used in the simulation has been increased to $0.055\mathrm{m}$ from the distance of $0.04\mathrm{m}$ used in the preliminary evaluation. This is done in order to increase the resolution of the TDoA. The potential increase in spurious peaks due to spatial aliasing can be compensated with an increase in the clutter rate of the CBMeMBer filter.

There are three speech sources used in this simulation unlike the preliminary evaluations whereby only two synthetic male and female speech signals were used. The speech signals comprise of two clean male speech signals and a clean female speech signal sampled at $16\mathrm{kHz}$. These speech signals are active at different time frames and

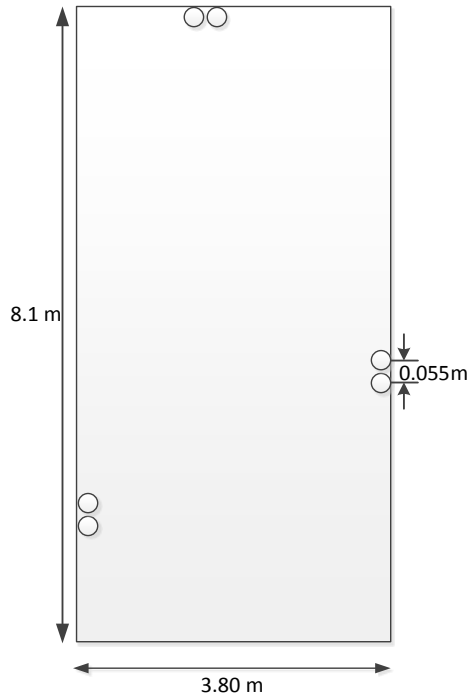


Figure 6.4: Room dimension and setup

there is overlap between the active speech signals. The spectograms of the speakers are shown in figure 6.5 while the speech signal plots of the these speakers are shown in figure 6.6.

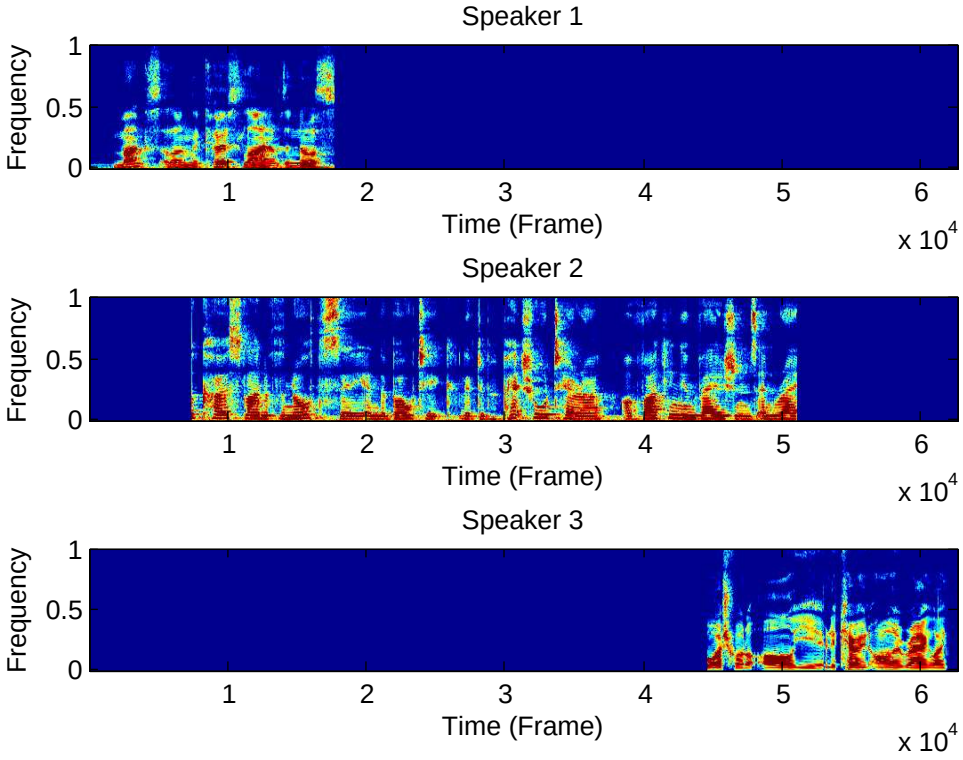


Figure 6.5: Spectogram of the three speech sources used in the simulation

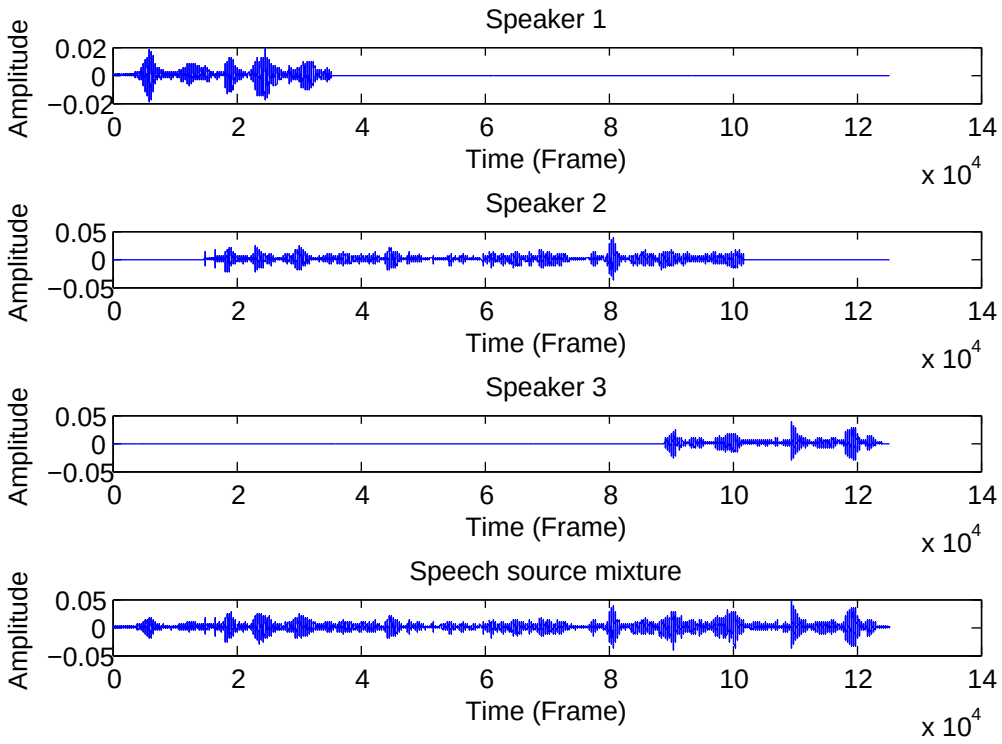


Figure 6.6: Plot of the three clean speech sources and the speech source mixture used in the simulation

Most of the settings in the CBMeMBer remain the same as the ones used in the preliminary evaluations in Section 6.3. The steady state velocity of, $\bar{v}$ and the rate constant, B of a speaker in the Langevin model remain the same at $1.4\mathrm{m/s}$ and $10\mathrm{Hz}$. The birth model is still a Gaussian spawned at the starting locations of the speakers but the covariance of the Gaussian differs according to the different noise or reverberation levels. The sigma, σ, of the normally distributed likelihood remains at 0.5% of the maximum TDoA between the sensor pair. The clutter rate, κ is set to 6 for the anechoic scenario but it is increased accordingly with the increase in noise level and reverberation time. The parameters in the track management algorithm has also been optimised accordingly. The number of allowable missed time frame for a tracked speech source has been reduced from 50 used in the preliminary evaluations to just 20 as this simulation has less silent gaps in the speech signals.

Using the room setup illustrated in figure 6.4, the effects of reverberation and noise on the tracking and source separation performance were evaluated. The room setup

was only changed when the effects of room size on performance of the proposed solution was studied. An anechoic room simulation which has no noise or reverberation was used as the control in the evaluations.

6.5 Effects of Reverberation

In order to study the effect of reverberation, the room setup as shown in figure 6.4 was simulated with varying reverberation time. The Signal to Noise Ratio (SNR) of the room is kept at 30dB in order to mitigate the effects of noise in the evaluation of reverberation time on the tracking and separation performance of the proposed solution. A simulation performed in a similar room setup with no reverberation or noise is used as a reference simulation. The results of the reference simulation are shown in figures 6.7 and 6.8.

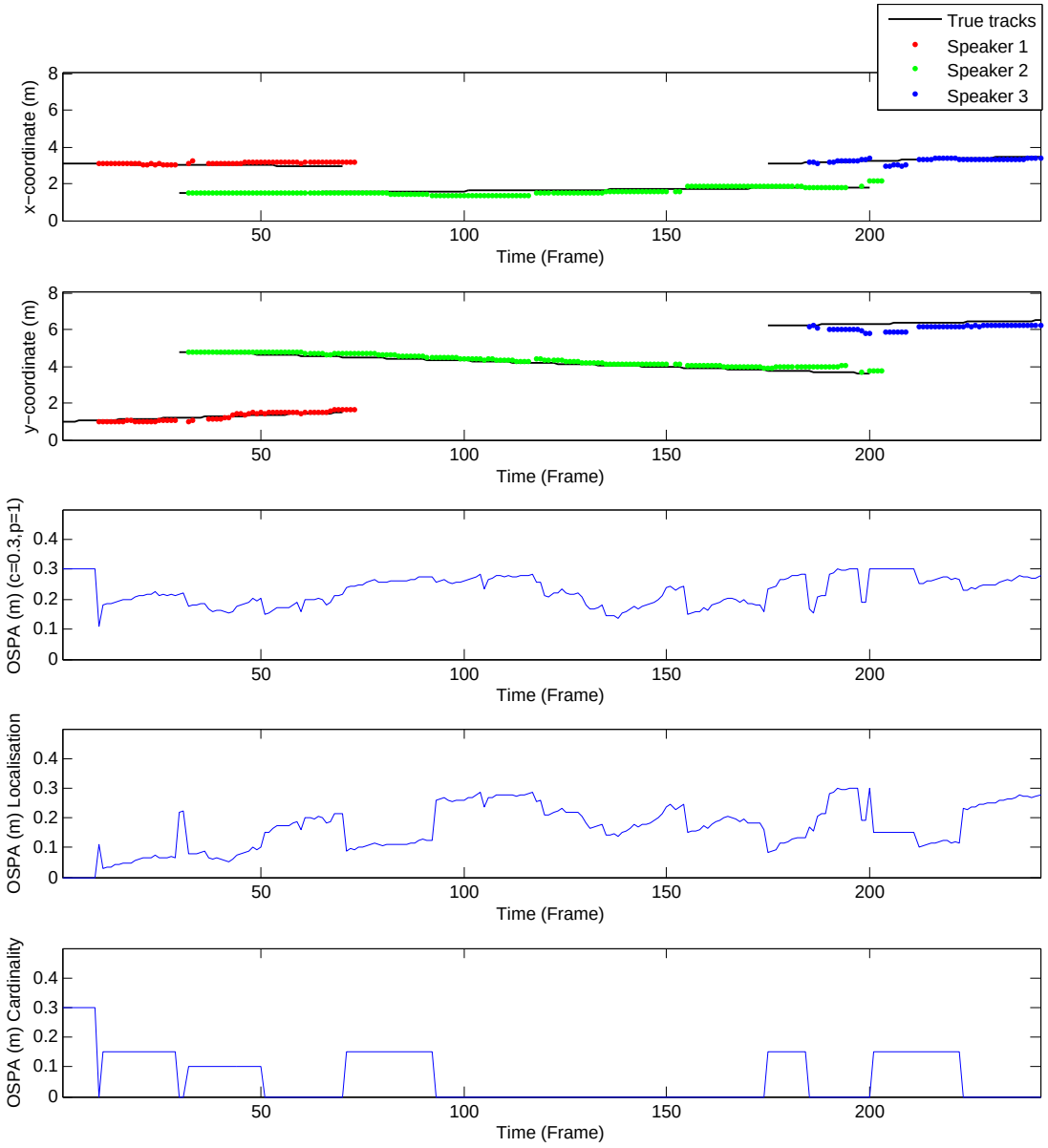


Figure 6.7: Tracking result of three speakers in an ideal scenario with no noise or reverberation

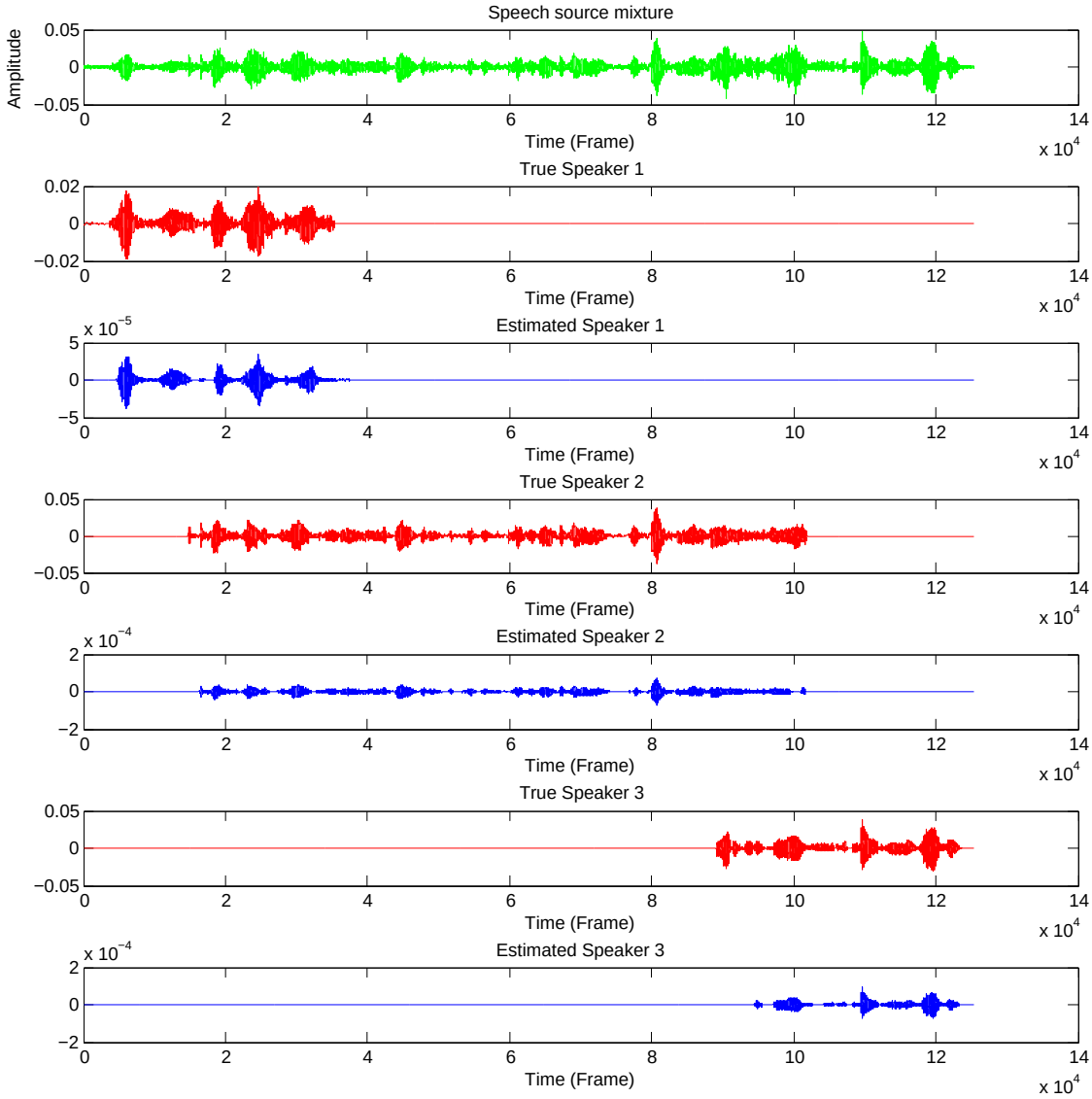


Figure 6.8: Plot of the three estimated speech sources in anechoic scenario with no noise or reverberation

Examples of the results of the effects of reverberation on the tracking performance are shown in figures 6.16, 6.10, and 6.11.

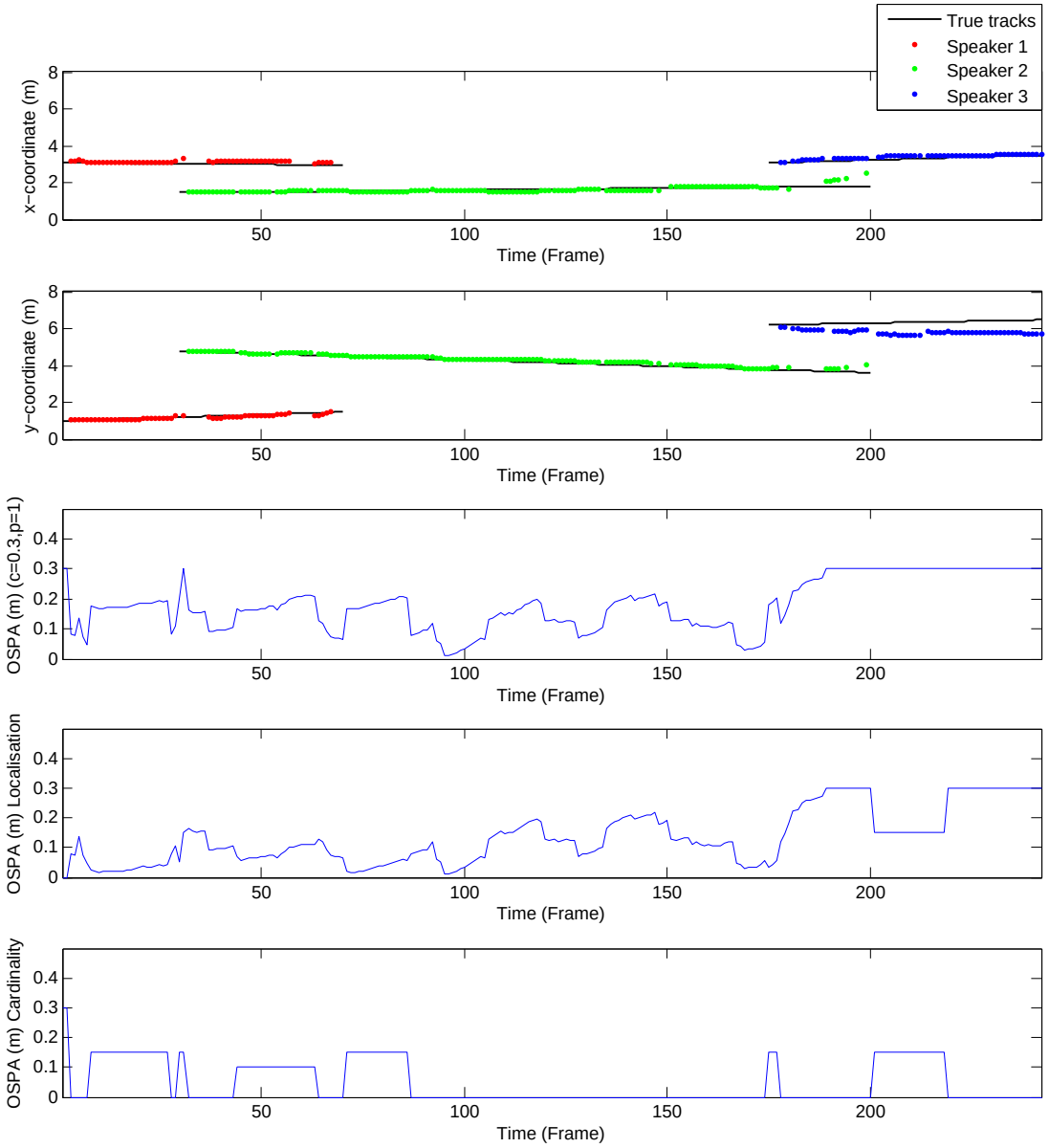


Figure 6.9: Tracking result of three speakers in a room with T60 reverberation time of 0.05s and SNR of 30dB

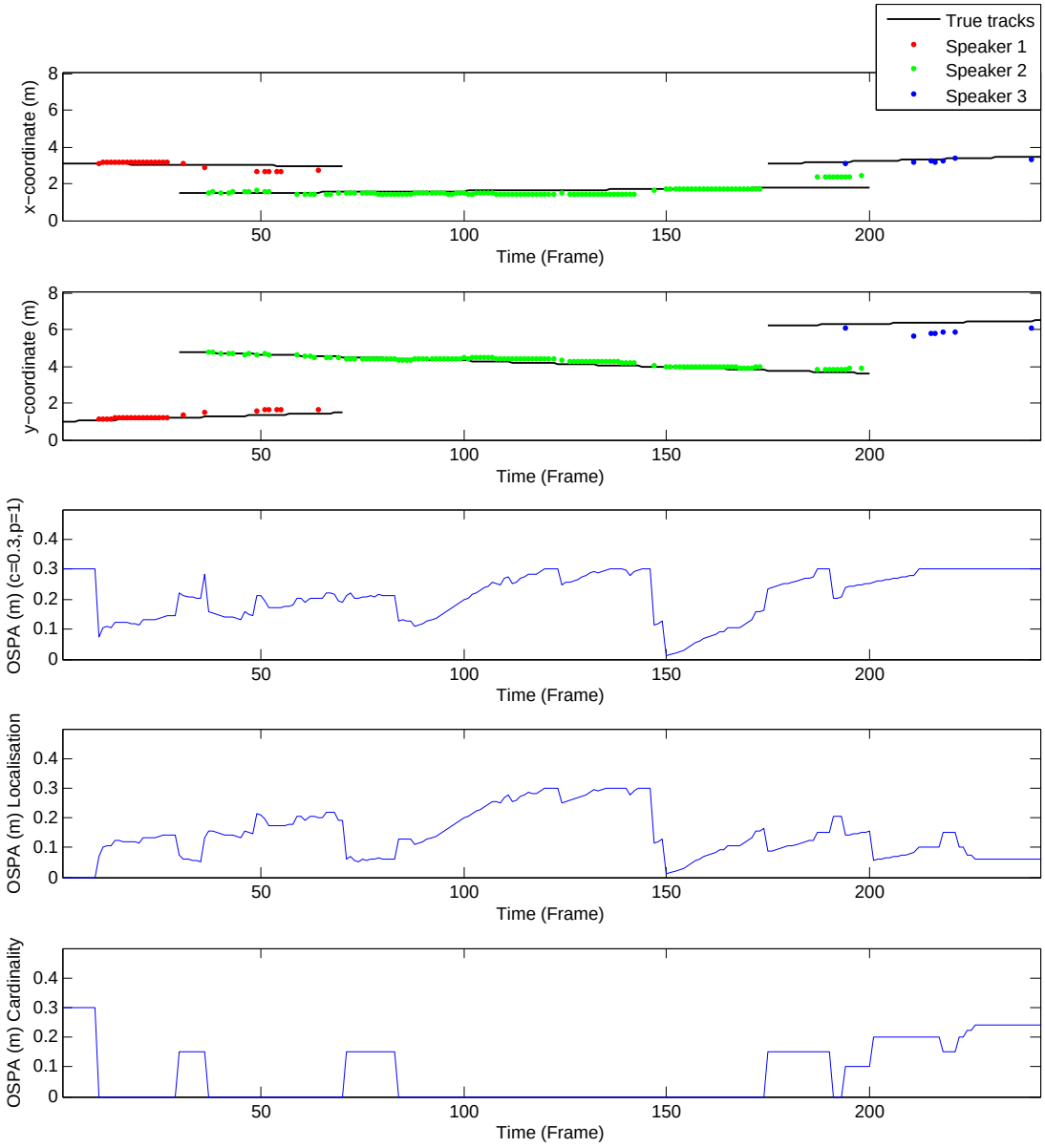


Figure 6.10: Tracking result of three speakers in a room with T60 reverberation time of 0.25s and SNR of 30dB

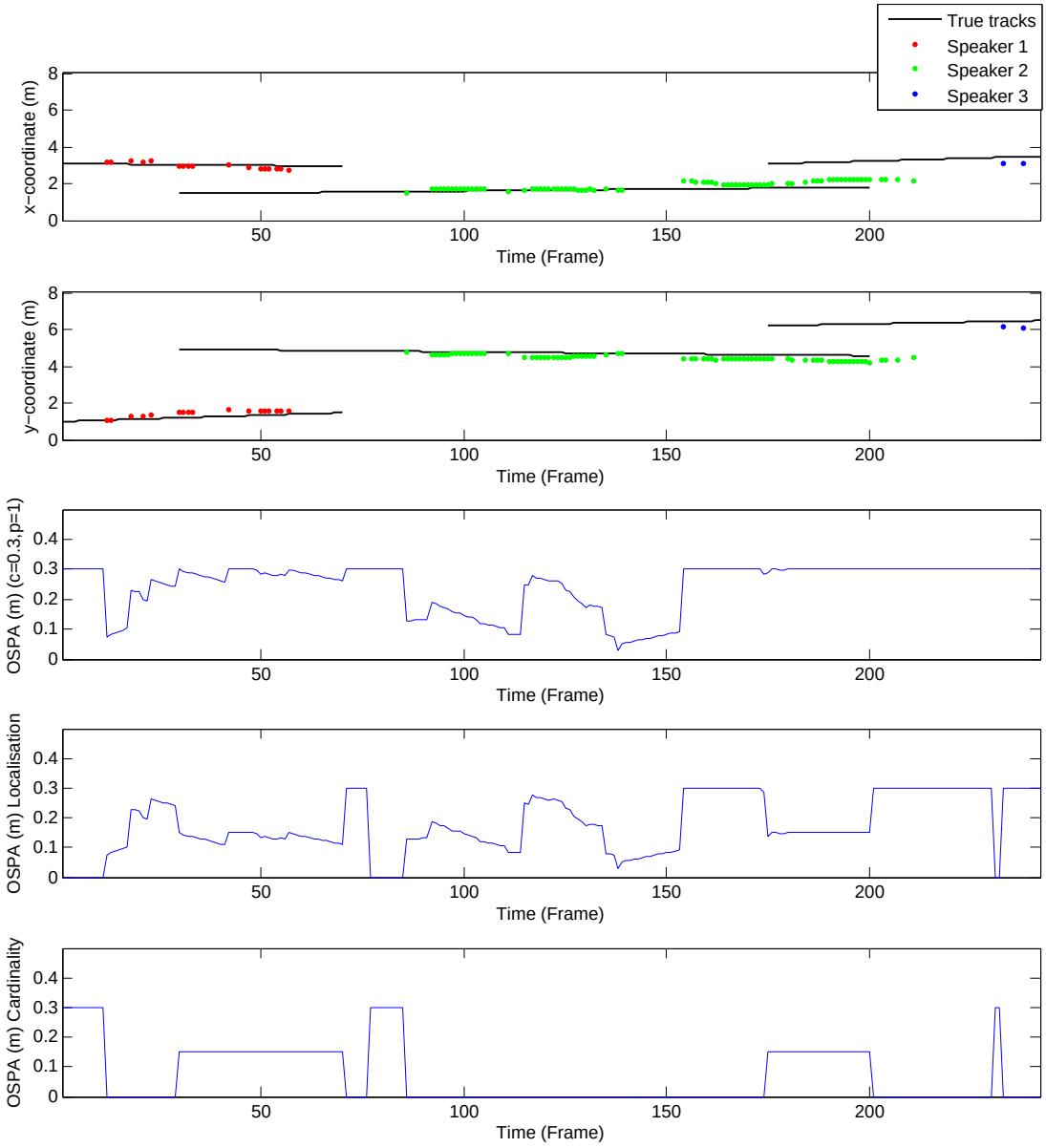


Figure 6.11: Tracking result of three speakers in a room with T60 reverberation time of 0.45s and SNR of 30dB

Examples of the effects of noise on the separation results are shown in figures 6.19, 6.13, and 6.14.

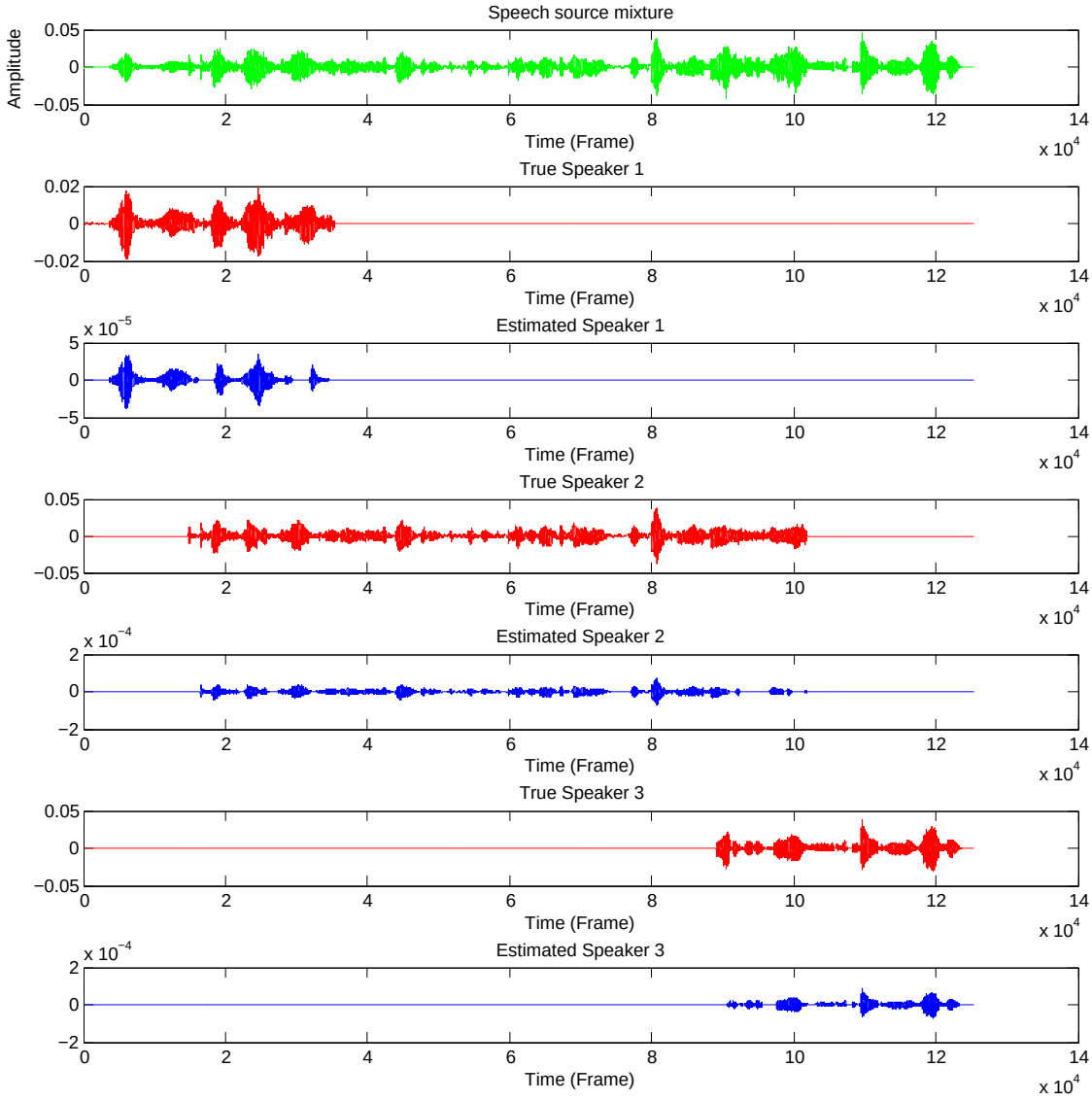


Figure 6.12: Estimated speech signals of the three speakers in room with T60 reverberation time of 0.05s and SNR of 30dB

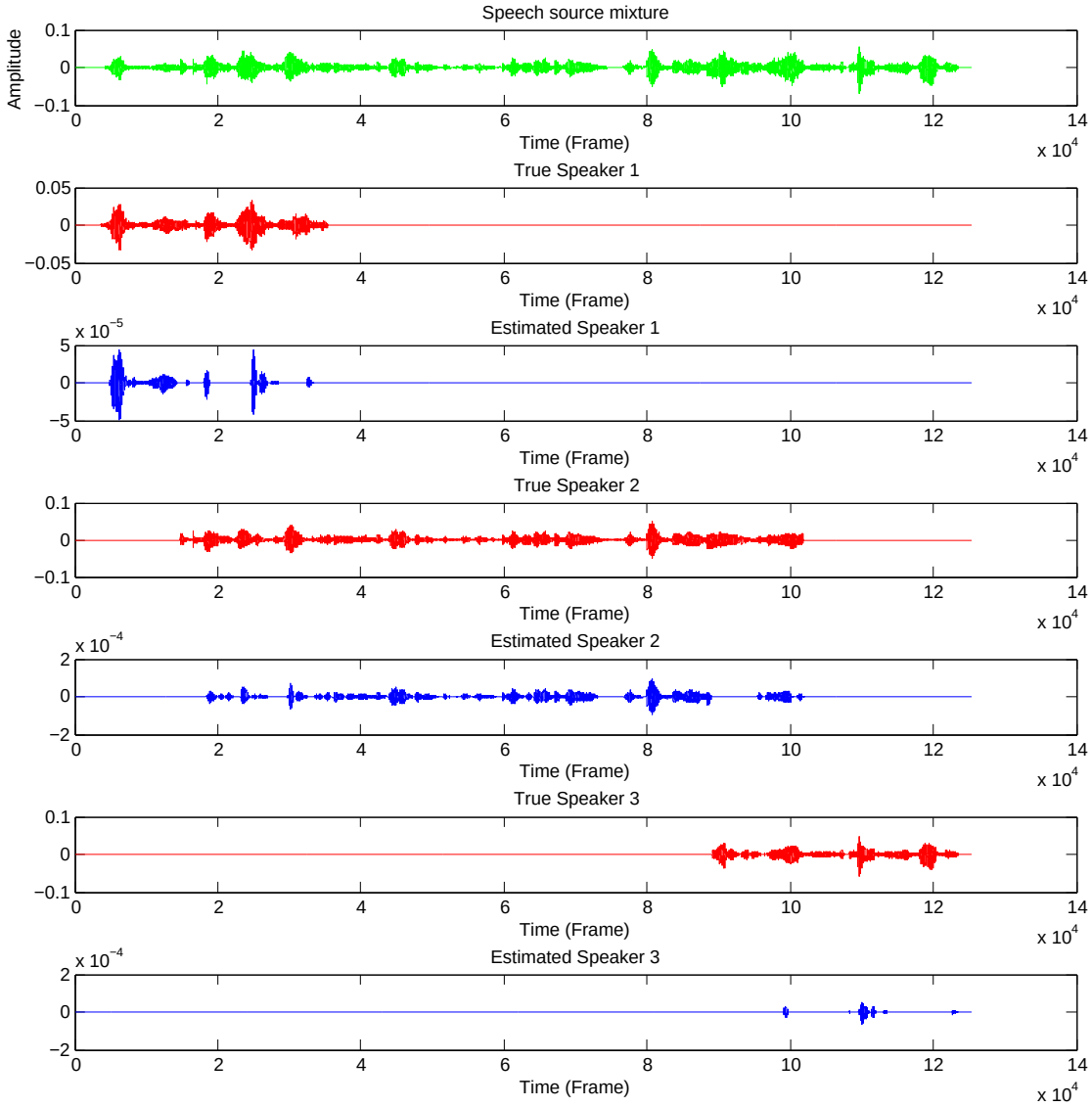


Figure 6.13: Estimated speech signals of the three speakers in room with T60 reverberation time of 0.25s and SNR of 30dB

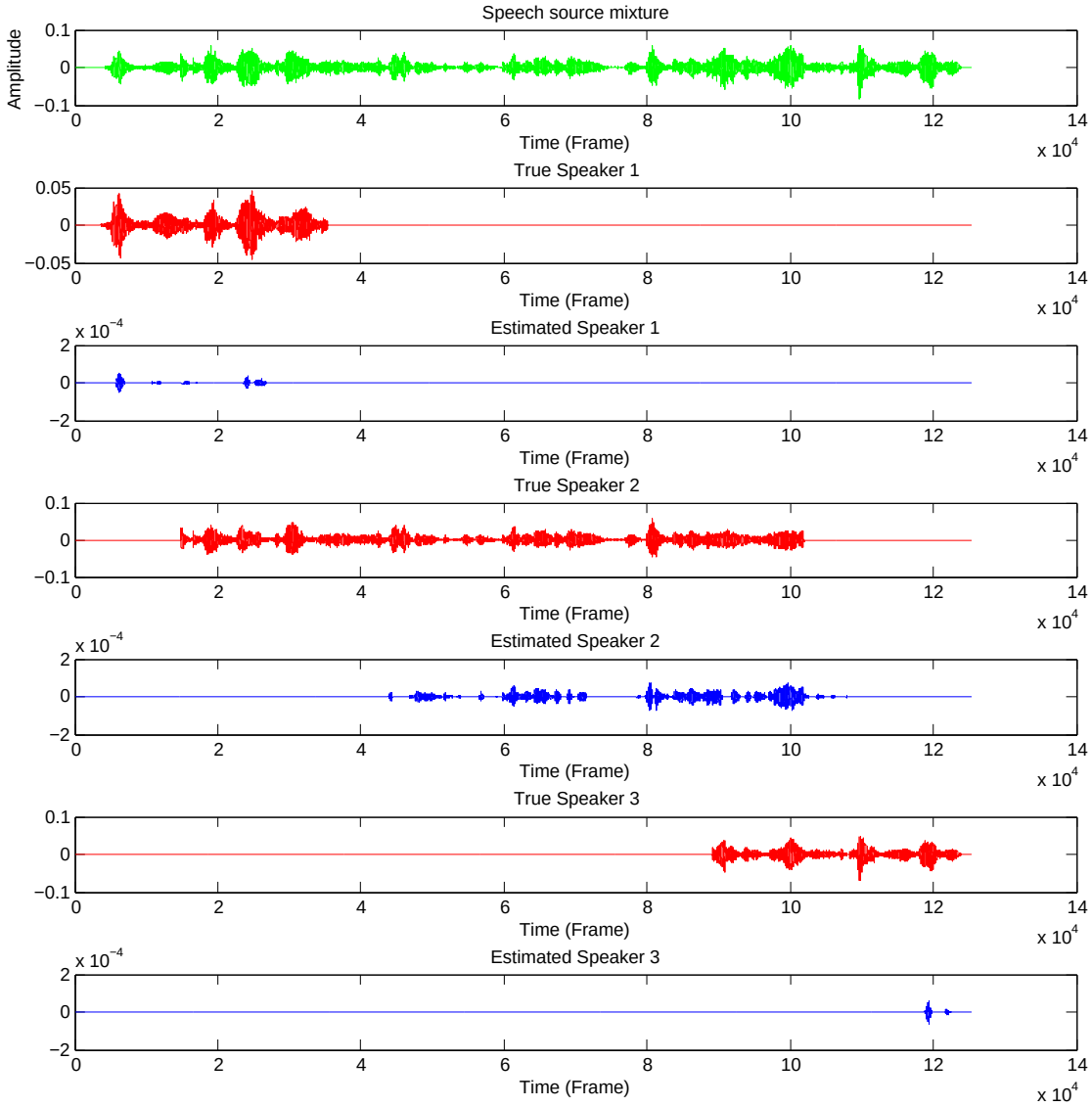


Figure 6.14: Estimated speech signals of the three speakers in room with T60 reverberation time of 0.45s and SNR of 30dB

The separation performance evaluated using the *BSS Toolkit* is summarised in figure 6.30.

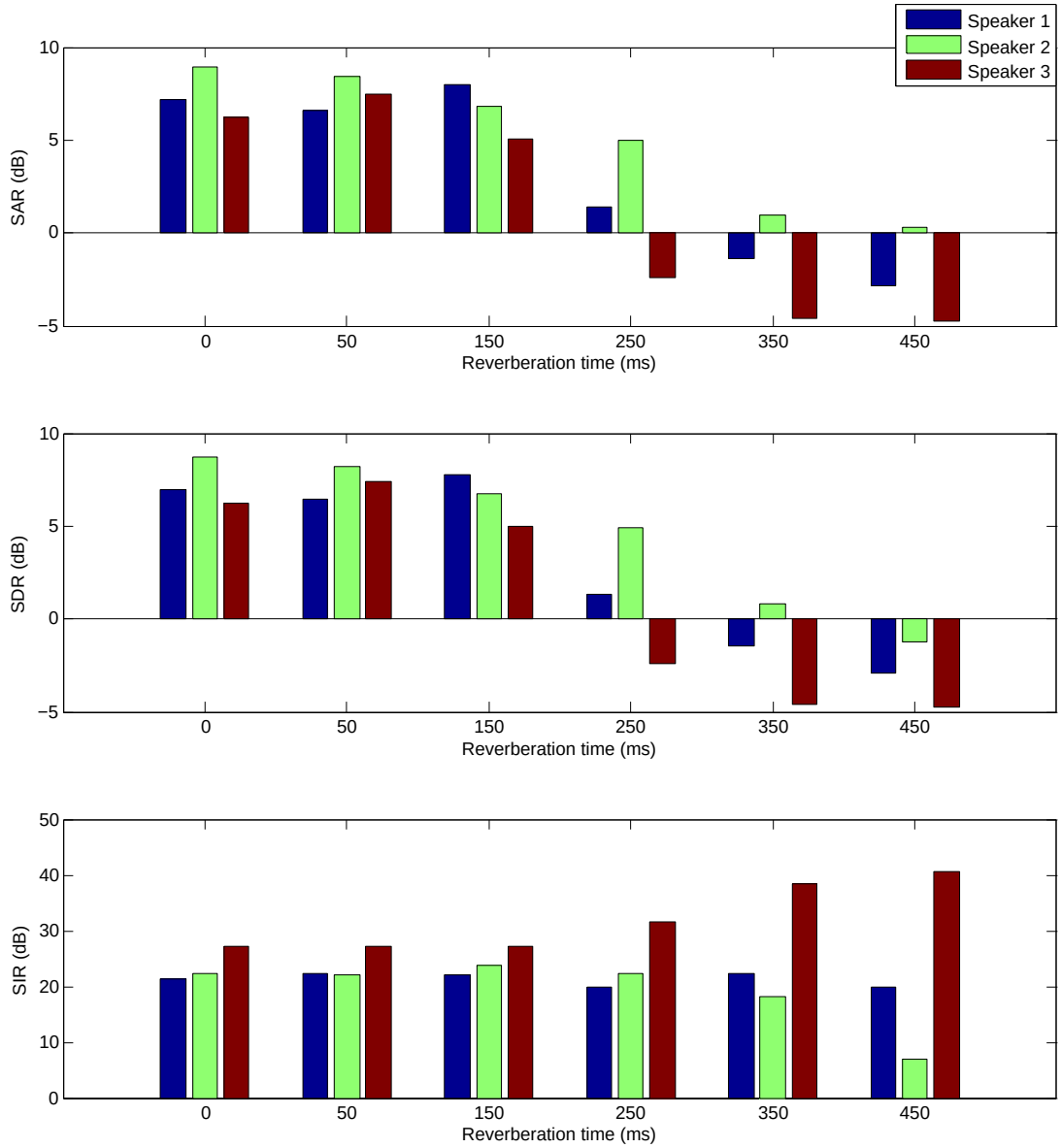


Figure 6.15: Separation performance in different room reverberations

6.5.1 Discussion

The results show that the proposed solution is capable of tracking and separating multiple moving speech sources in scenarios with various room reverberation. The tracking and separation performance is best when the room has low reverberation - as shown in figures 6.16 and 6.19. When there is no reverberation in the room, the proposed solution has an average per target error $0.2292\mathrm{m}$. With a low T60 reverberation time in the room of $0.05\mathrm{s}$, the proposed solution only has an average per target error of $0.1798\mathrm{m}$. As can be observed from the general trend shown in the results, both the tracking and separation performance deteriorate as the room reverberation increases.

The proposed solution loses track of the speech sources when they overlap in rooms with high reverberation as illustrated in figure 6.10 and 6.11. The proposed solution loses part of the tracked speech sources starting when the T60 room reverberation is 0.25s. When T60 room reverberation has increased to 0.45s, the average per target error has also increased to 0.2444m. The deconstructed OSPA shows that the deterioration in the performance is mainly contributed by the cardinality error component due to the missed speech sources as the localisation error component is fairly low when the speech sources are tracked.

The separation performance is closely associated with the tracking performance. This is mainly due to the construction of the TF mask. The TF mask used to separate the speech sources relies on the position and identity information of the tracked speech sources to obtain the associated acoustic feature. Hence, the separation performance is affected if the estimated acoustic features used to construct the TF masks are not accurate. When the state estimates are accurate, the separation performance of the proposed solution is good too. When there is no reverberation in the room, the average SAR, SDR and SIR of the three speakers are 7.4528dB, 7.3027dB and 23.6408dB. In the room with $T60 = 0.05\text{s}$, the average SAR, SDR and SIR of the three speakers are 7.4811dB, 7.3487dB and 23.8939dB. At the highest room reverberation of $T60 = 0.45\text{s}$, the average SAR, SDR and SIR of the three speakers are -2.4305dB, -2.9594dB and 22.5151dB. The separation performance deteriorates as the room reverberation increases as illustrated in the bar chart shown in figure 6.30. The decrease in the SAR and SDR values are obvious but the SIR values defy the general trend. This is due to the ways in which SIR and $e_{\text{int}}(t)$ are calculated as shown in equations (6.9) and (6.5). When the proposed solution loses track of the speech source, the portion of the missed speech source will be masked and not be reconstructed. As the estimated speech source only accounts for parts of the speech which gets tracked, $e_{\text{int}}(t)$ will be a small value. $e_{\text{int}}(t)^2$ is the denominator in the calculation of the SIR. With a smaller value of $e_{\text{int}}(t)$, the value of SIR increases. Hence, SIR is not really a good measure of separation for the proposed solution.

The deterioration in both the tracking and separation performance can be attributed to assumption mismatch. The assumption mismatch happens in the acoustic feature extraction stage and also the multi-target tracking stage. As shown in equation (5.3),

the acoustic transfer function used in the proposed solution is

$$\mathcal{H}_{21} = \frac{Y_2(\tau,\omega)Y_1^*(\tau,\omega)}{Y_1(\tau,\omega)Y_1^*(\tau,\omega)}. \tag{6.12}$$

This equation is actually an approximation of

$$\mathcal{H}_{21} = \frac{Y_2(\tau,\omega)Y_1^*(\tau,\omega)}{Y_1(\tau,\omega)Y_1^*(\tau,\omega)} e_{\text{reverb}}(\tau,\omega) \tag{6.13}$$

whereby e_{reverb} is the multiplicative reverberation error term. According to Dvorkin [116], acoustic transfer function in equation (6.13) can be approximated to the one shown in equation (5.3)when the reverberation is low as $e_{\text{reverb}} \approx 1$. As the reverberation time increases, the effect of e_{reverb} on the acoustic transfer function is more prominent so the acoustic features extracted using $\mathcal{H}_{21}$ is no longer as accurate as those extracted in the low reverberation scenario. The *W-disjoint orthogonality* assumption is also violated when the reverberation in the room is high. The assumption that only one speech source is active at each time frequency is no longer valid as the reflections of these active speech sources will also appear in the same time frequency domain. The assumption mismatch with regards to reverberation in the acoustic feature extraction stage results in more spurious peaks getting extracted and used as observations in the CBMeMBer filter.

As explained earlier in Section 4.2, the CBMeMBer filter works best in scenarios with high detection rate and low clutter. The increase in reverberation time will result in more clutter, which would affect the performance of the CBMeMBer filter. Furthermore, the CBMeMBer works under the assumption that the observed measurements and clutter are independent. In the "conference room scenario", the reverberation due to the speech sources are not really independent. The reflections of the speech sources also move accordingly as the speech sources move across the room. When the room reverberation is low such as $T60 = 0.05\text{s}$, the clutter due to reverberation can still be assumed to be independent. However as the T60 room reverberation time increases 0.25s, the level of reverberation becomes much more significant. The spurious TDoA peaks resulting from reverberation follow a pattern in their movements. Thus, clutter which are contributed by these spurious peaks will occasionally get confirmed as a track belonging to an active speech source.

6.6 Effects of Noise

Apart from the studying the effects of reverberation on the tracking and separation performance, the effects of noise on the proposed solution's performance was investigated as well. In order to rule out the effect of reverberation on the results, the room reverberation for all the different SNR simulations is set at $T60 = 0.05\text{s}$. The 6 different SNR used in the simulations are 20dB, 22dB, 24dB, 26dB, 28dB and 30dB. The noise used is an additive white Gaussian noise. Examples of the tracking results are shown in figures 6.16, 6.17, and 6.18.

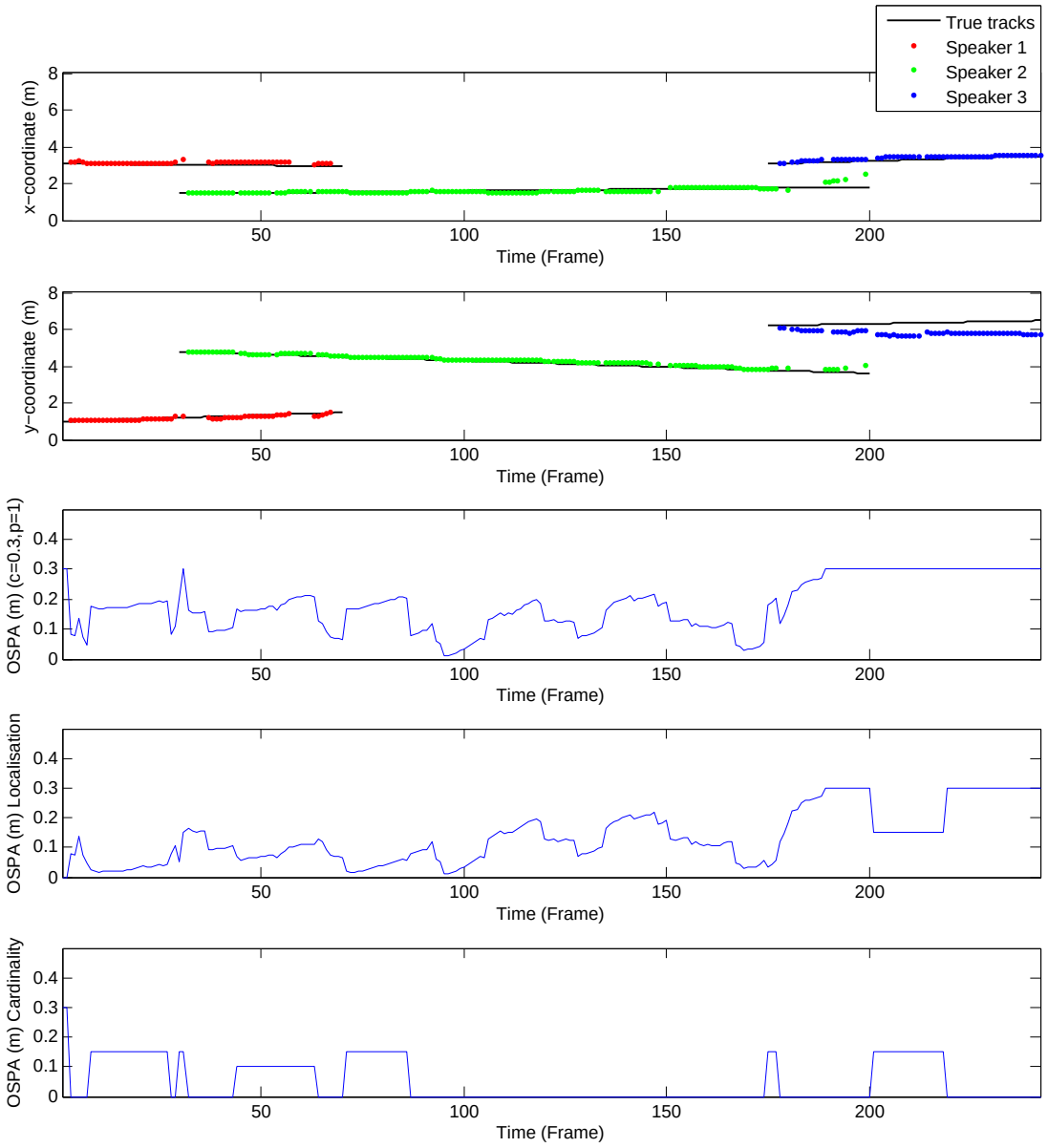


Figure 6.16: Tracking result of three speakers in a room with T60 reverberation time of 0.05s and SNR of 30dB

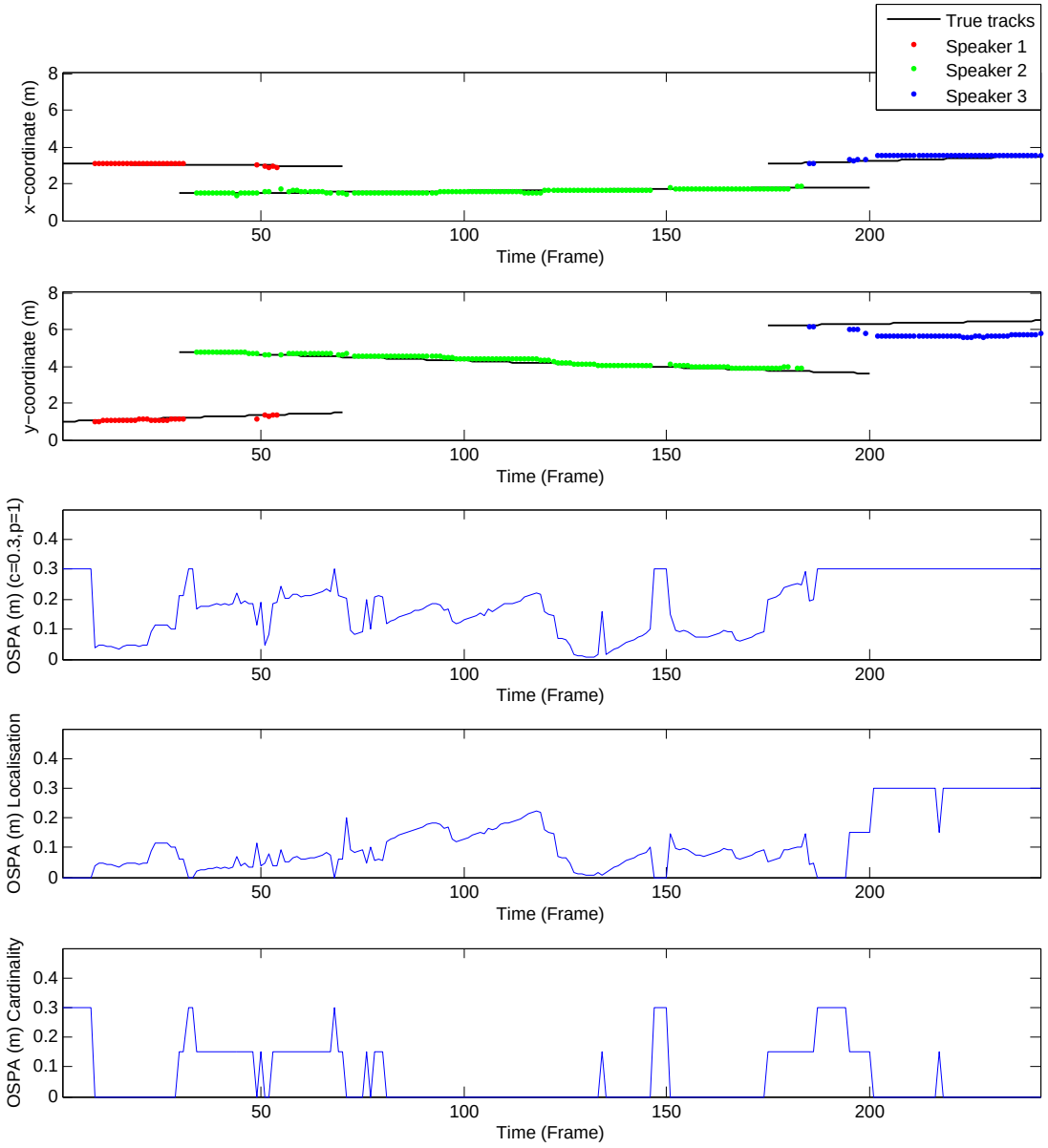


Figure 6.17: Tracking result of three speakers in a room with T60 reverberation time of 0.05s and SNR of 26dB

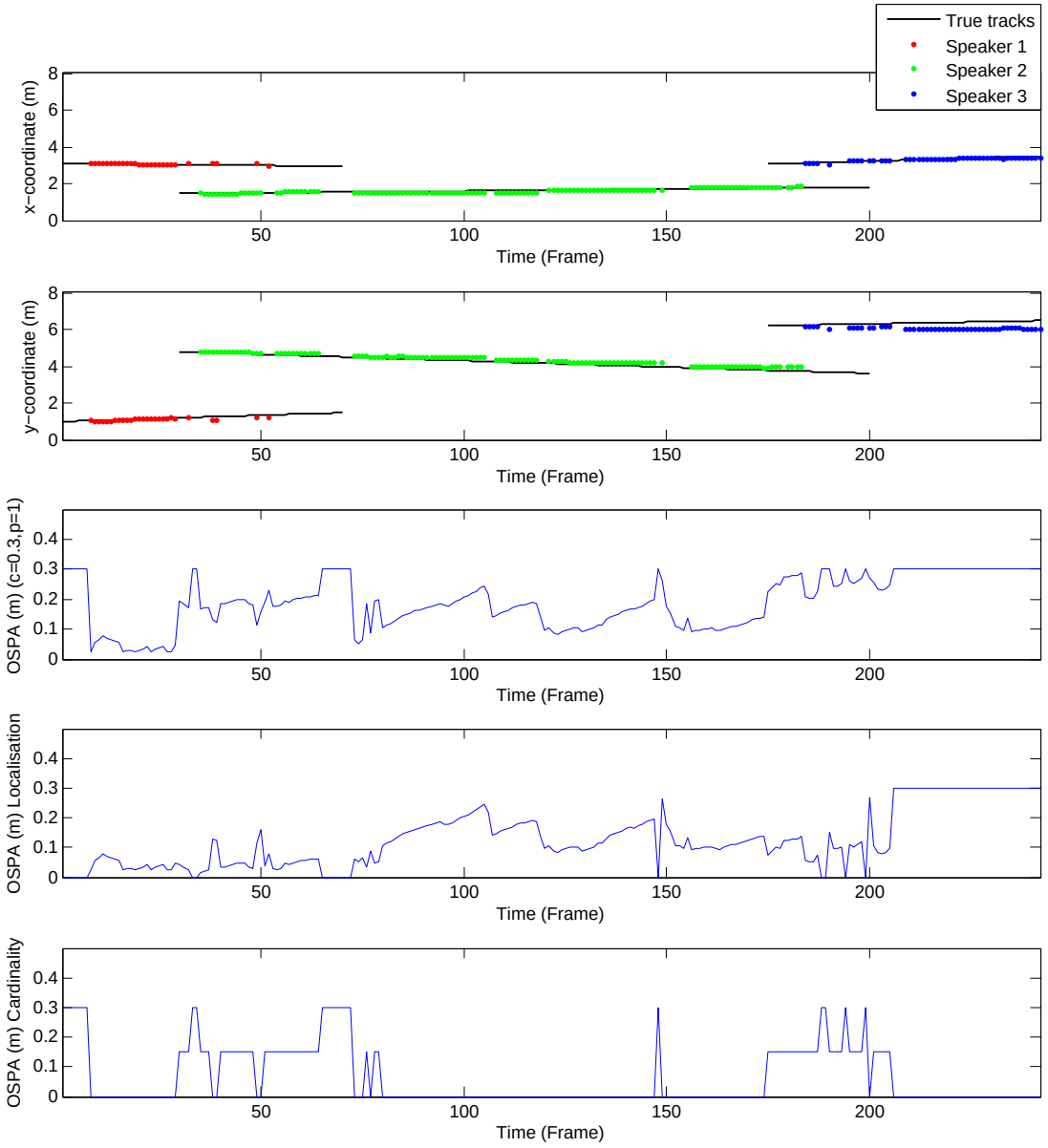


Figure 6.18: Tracking result of three speakers in a room with T60 reverberation time of 0.05s and SNR of 20dB

The effects of noise on the separation results are shown in figures 6.19, 6.20, and 6.21. The results of the reference simulation are shown in figures 6.7 and 6.8.

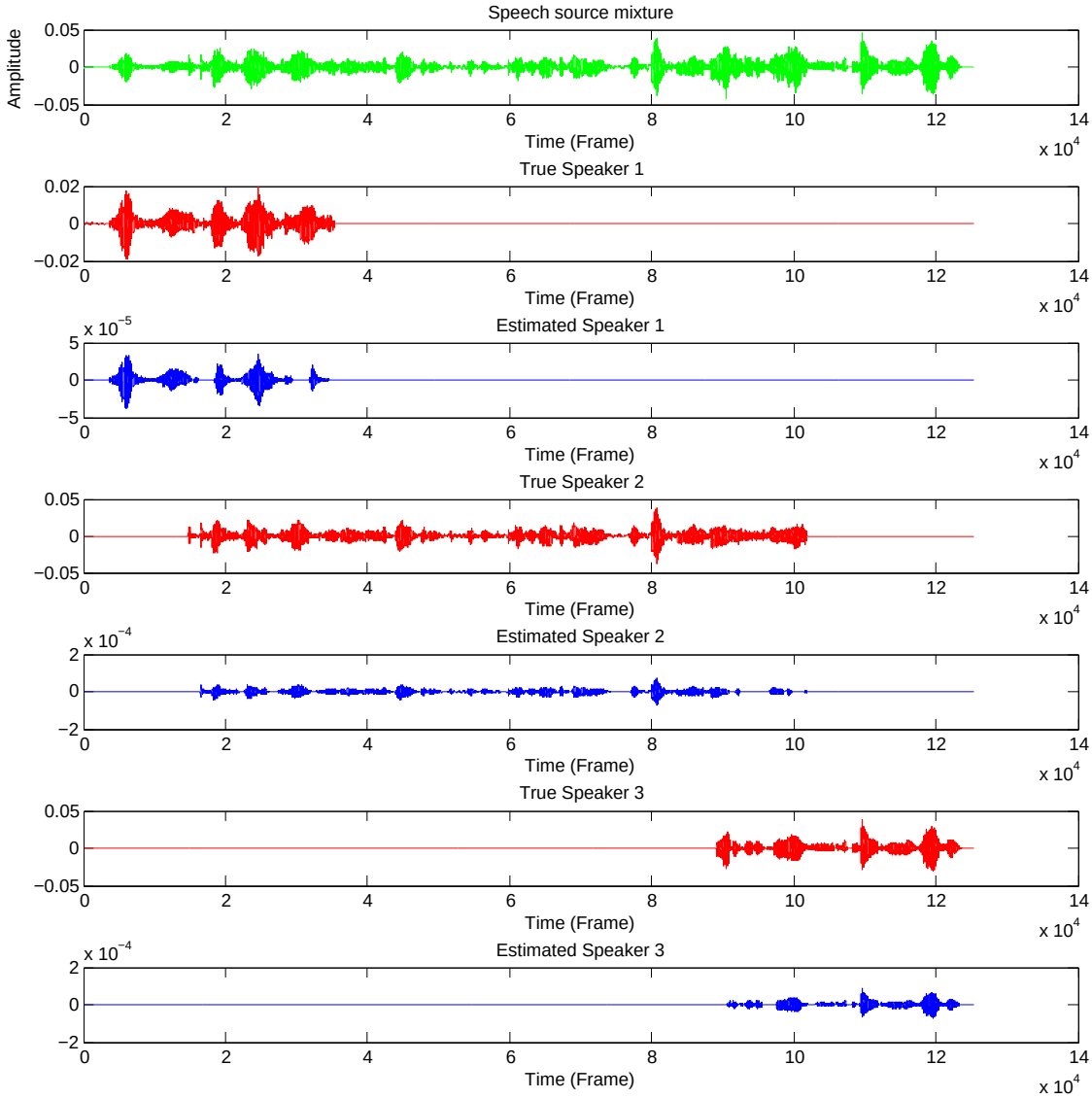


Figure 6.19: Estimated speech signals of the three speakers in room with T60 reverberation time of 0.05s and SNR of 30dB

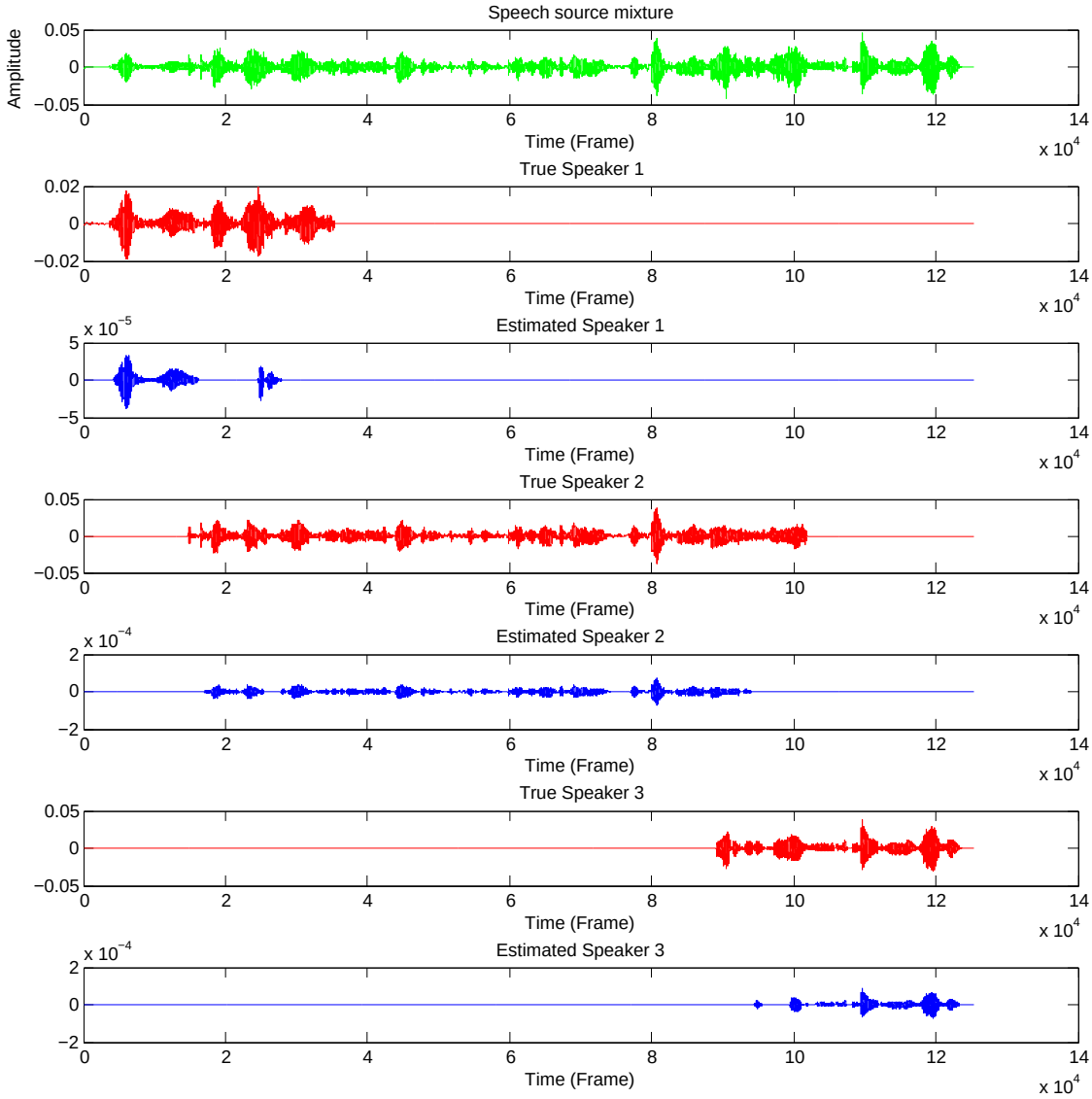


Figure 6.20: Estimated speech signals of the three speakers in room with T60 reverberation time of 0.05s and SNR of 26dB

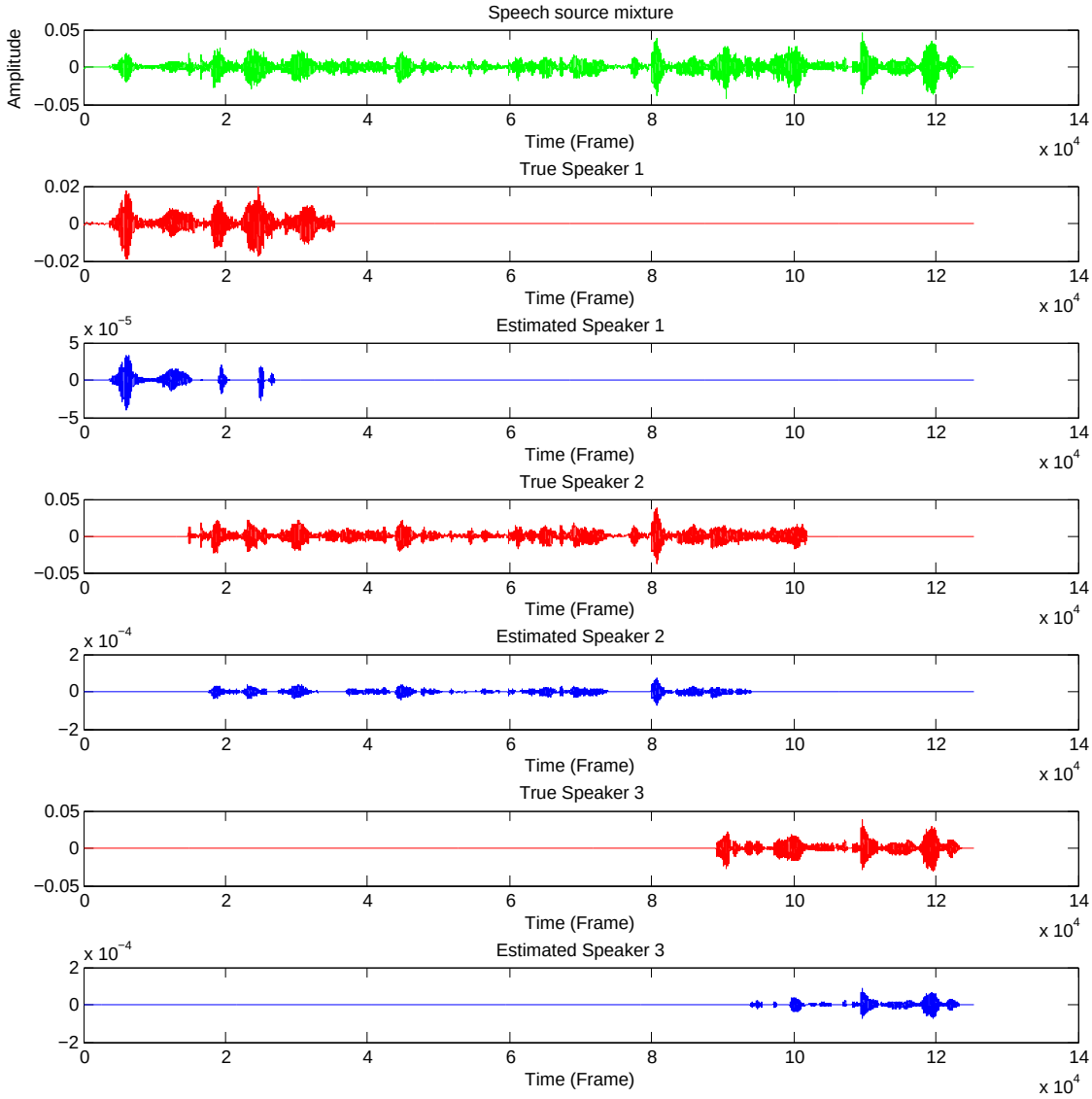


Figure 6.21: Estimated speech signals of the three speakers in room with T60 reverberation time of 0.05s and SNR of 20dB

The separation performance evaluated using the *BSS Toolkit* is summarised in figure 6.22.

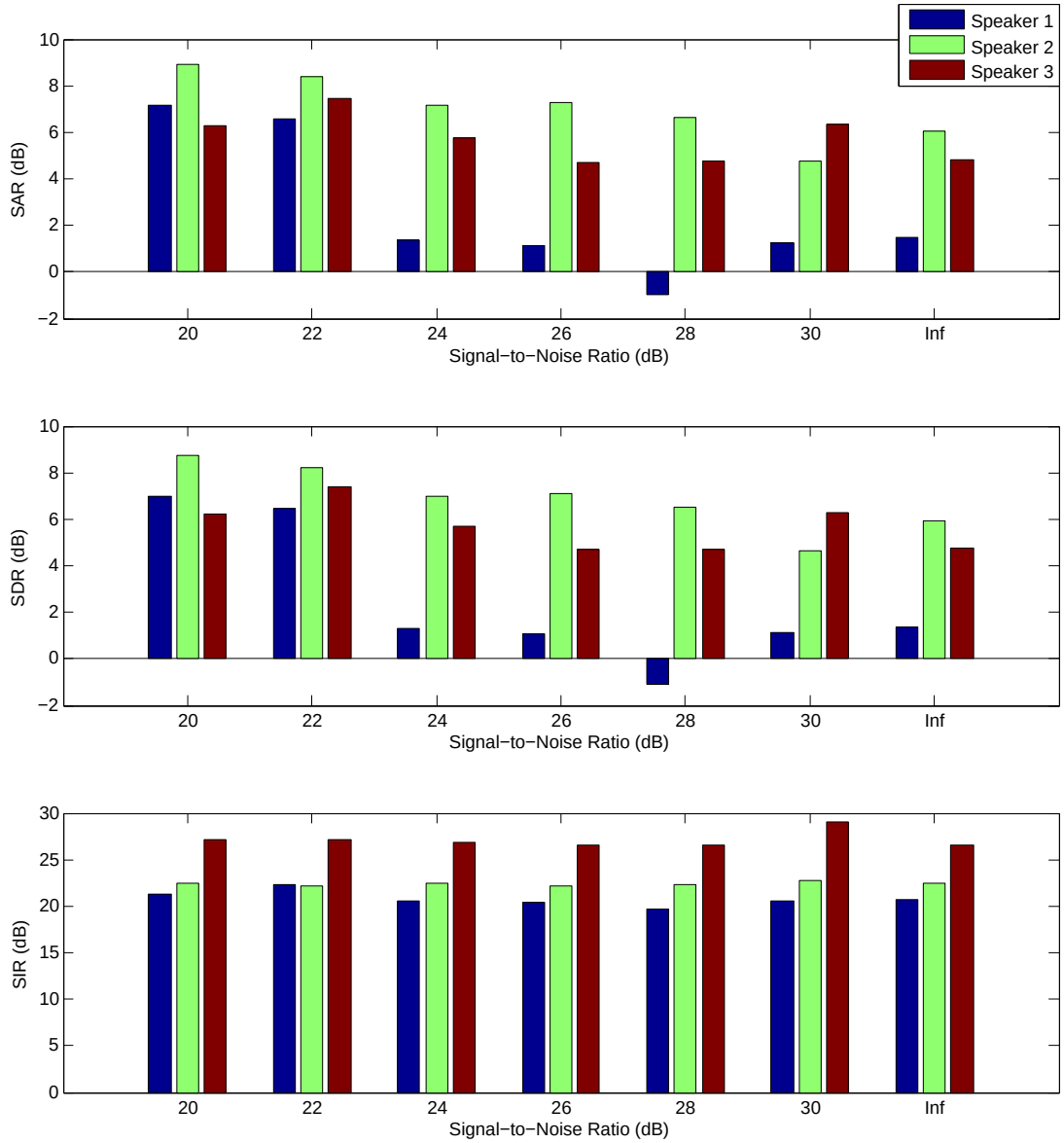


Figure 6.22: Separation performance in different Signal to Noise Ratios (SNR)

6.6.1 Discussion

The effect of noise on the tracking and separation performance is as expected. The trend in the results shows deterioration in the tracking performance as the speech source mixture's SNR decreases. The trend can be clearly observed from figures 6.16 to 6.18 as the proposed solution slowly loses track of the speech sources when they overlap. At the highest SNR of $30\mathrm{dB}$, the proposed solution produces a per target error of $0.1798\mathrm{m}$. The performance degrades to an average per target error of only $0.1901\mathrm{m}$ when the SNR has decreased to $20\mathrm{dB}$. The tracking performance of the proposed solution is $0.1827\mathrm{m}$ when there is moderate noise levels in the room and the

SNR of the received speech mixture is 26dB. The main contribution to the OSPA error when the speech sources overlap in time is due to the error in estimating the number of speech sources as the cardinality error component has a maximum error of 0.3m.

In terms of separation performance, the performance of the separation algorithm also suffers as the SNR of the speech mixture decreases. This is illustrated in figures 6.19, 6.20, and 6.21 as less of the speech sources are reconstructed. The same trend is observed in the SAR and SDR readings in figure 6.22. The SIR remains largely stationary throughout the change in the SNR of the speech source signals. The highest SNR of 30dB yields an average SAR, SDR and SIR of 7.4811dB, 7.3487dB and 23.8939dB while the lowest SNR of 20dB has an average SAR, SDR and SIR of 4.1065dB, 4.0238dB and 23.2551dB.

Despite the deterioration in the tracking and separation performance as the SNR of the speech mixture decreases, this degradation in performance is less than the degradation in performance due to reverberation. In terms of tracking performance, the difference between the OSPA distance of signals the highest SNR and the lowest SNR is 0.0103m but the difference between the OSPA distance of the rooms with the lowest and highest reverberation time is 0.0646m. This can be attributed to the independent white noise matching the modelling assumptions in the the tracking algorithm. The CBMeMBer assumes that the received measurements are independent and uncorrelated. As this assumption is met, the effects of noise on the performance of the proposed solution is mitigated. Furthermore, the spurious peaks due to noise do not follow any movement patterns so dynamics model will not mistake these measurements for that of a moving speech source.

As can be observed from the results, the spikes in the OSPA distance occurs when the speech sources overlap. This is mainly caused by the extra spurious peaks which are observed. When the SNR decreases, the number of these spurious peaks increase as well. As the CBMeMBer is unable to determine if the extra observations are noise or spurious peaks, it loses track of the speech sources. The lost speech sources result in an increase in the cardinality error component of the OSPA distance. When the proposed solution loses track of the speech source, the portion of the missed speech source will be masked and not be reconstructed. As stated earlier in Section 6.5.1, the SIR remains stationary despite the changes in the SNR mainly because of the way

SIR and $e_{\text{int}}(t)$ are calculated. The method of SIR and $e_{\text{int}}(t)$ is also the reason the cardinality error has little effect on the SIR. The SIR also does not change in value even when the cardinality error increases. Equations (6.9) and (6.5) shows the formulae used to calculate SIR and $e_{\text{int}}(t)$. A projection of the speech source is used to calculate the SIR. Such a measure would work when the assumption that the speech sources are fully reconstructed is met. In this "conference room problem", the tracking algorithm might lose track of the speech sources if the reverberation or noise levels are too high. Under such circumstances, there is not much of the speech source to be projected and $e_{\text{int}}(t)$ which is calculated from the projected speech source will have a small value. The small value of $e_{\text{int}}(t)$ results in a higher value of the SIR.

6.7 Effects of Room Size

Apart from the effects of reverberation and noise level, the effects of room sizes on the performance of the proposed solution is also investigated. The three room sizes can be categorised as small, medium and large. The small room which has a dimension of $3.0\text{m} \times 3.0\text{m}$ corresponds to a typical small office. The medium sized room used which is $8.1\text{m} \times 3.8\text{m}$ corresponds to a common conference room. The large room which represents a typical lecture theater has a dimension of $15\text{m} \times 20\text{m}$. Although the proposed solution is meant to be applied in a "conference room scenario", the effects of which the different room sizes have on the performance will allow the algorithm to be optimised for those scenarios. In order to eliminate reverberation time and noise levels as variables in this experiment, all three rooms were simulated with a reverberation time of $T60 = 0.15\text{s}$ and a SNR of 30dB. The tracking results are shown in figures 6.23, 6.24 and 6.25 while the separation results are shown in figures 6.26, 6.27 and 6.28.

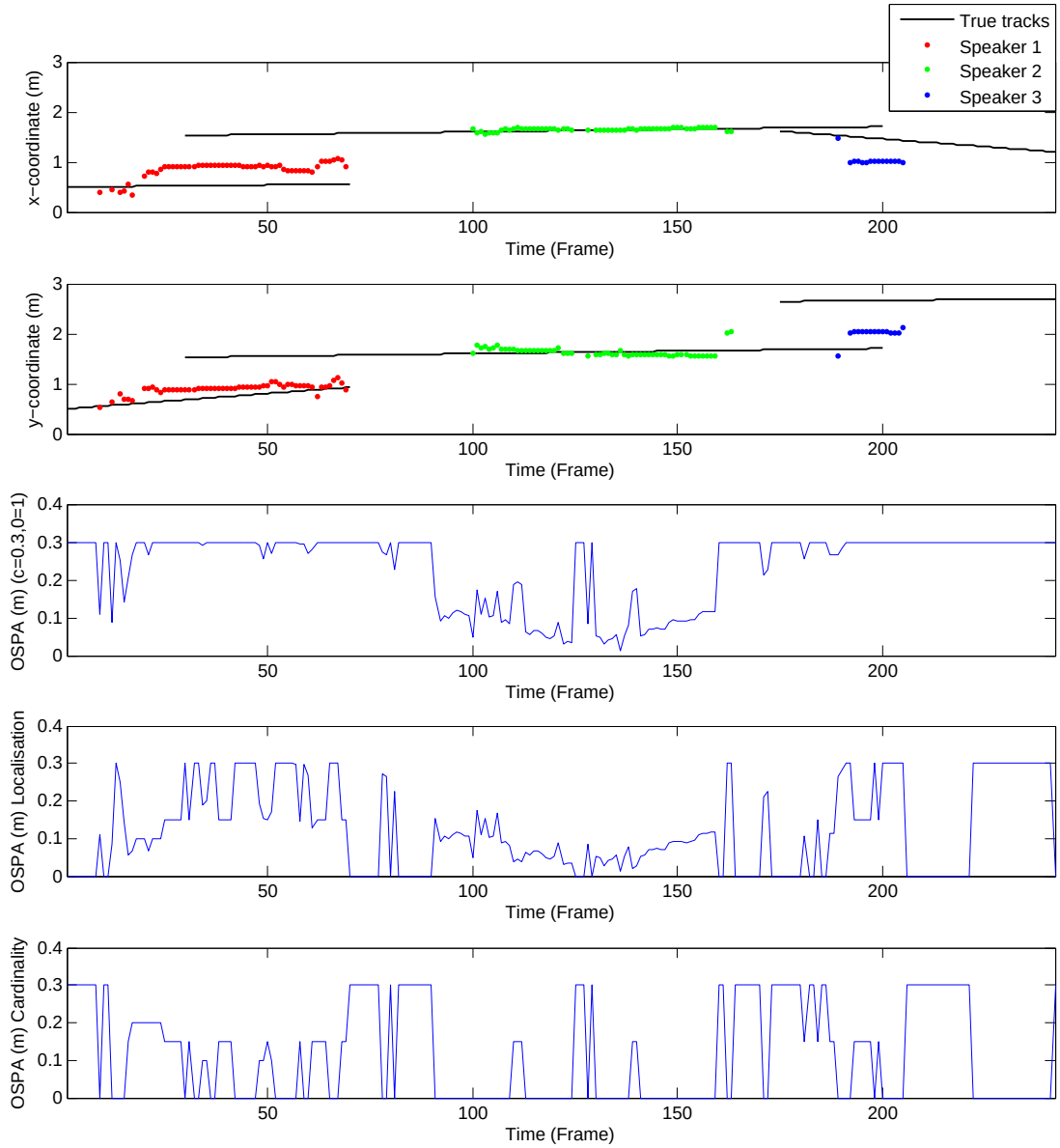


Figure 6.23: Tracking result of three speakers in a small room

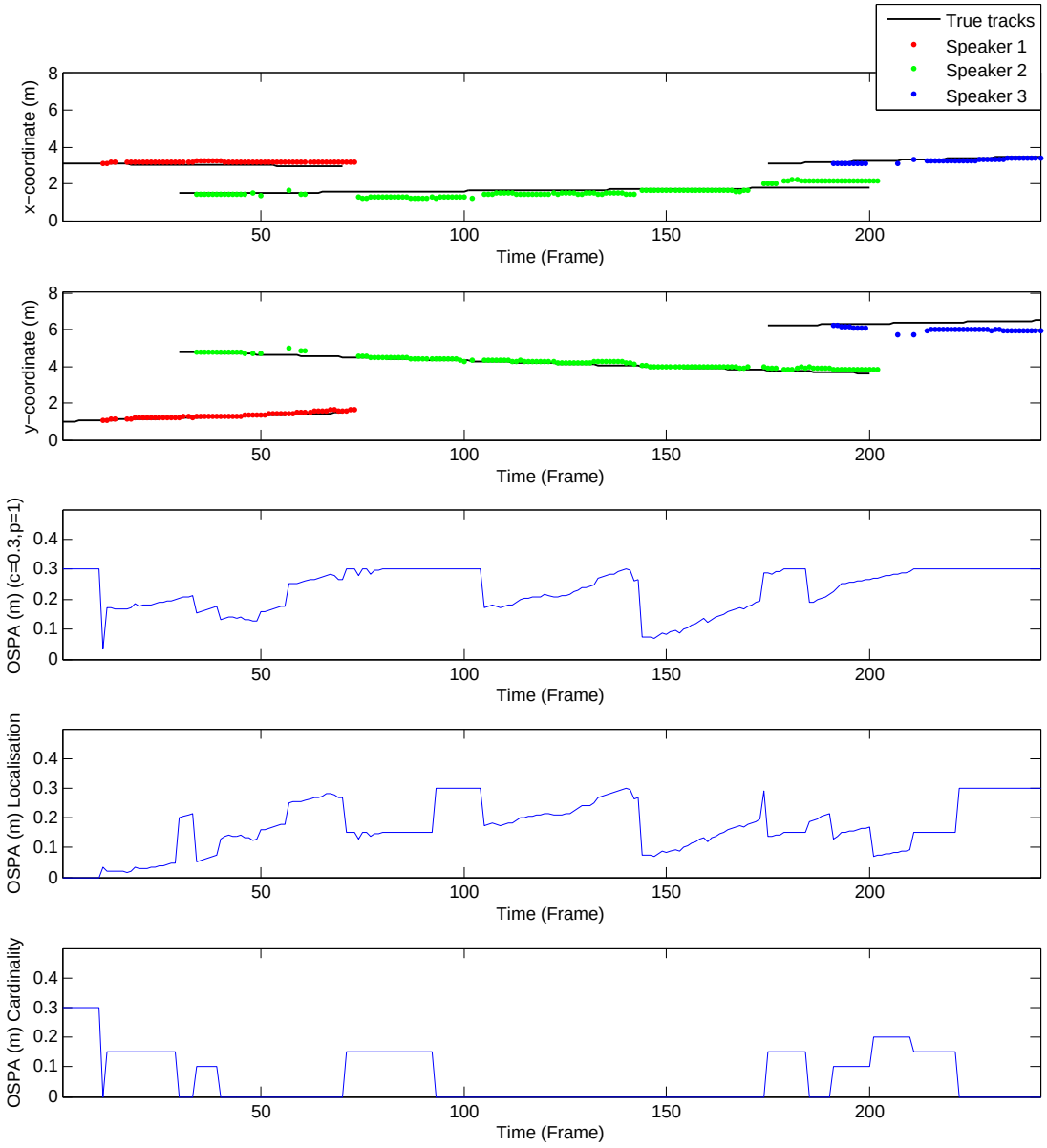


Figure 6.24: Tracking result of three speakers in a medium room

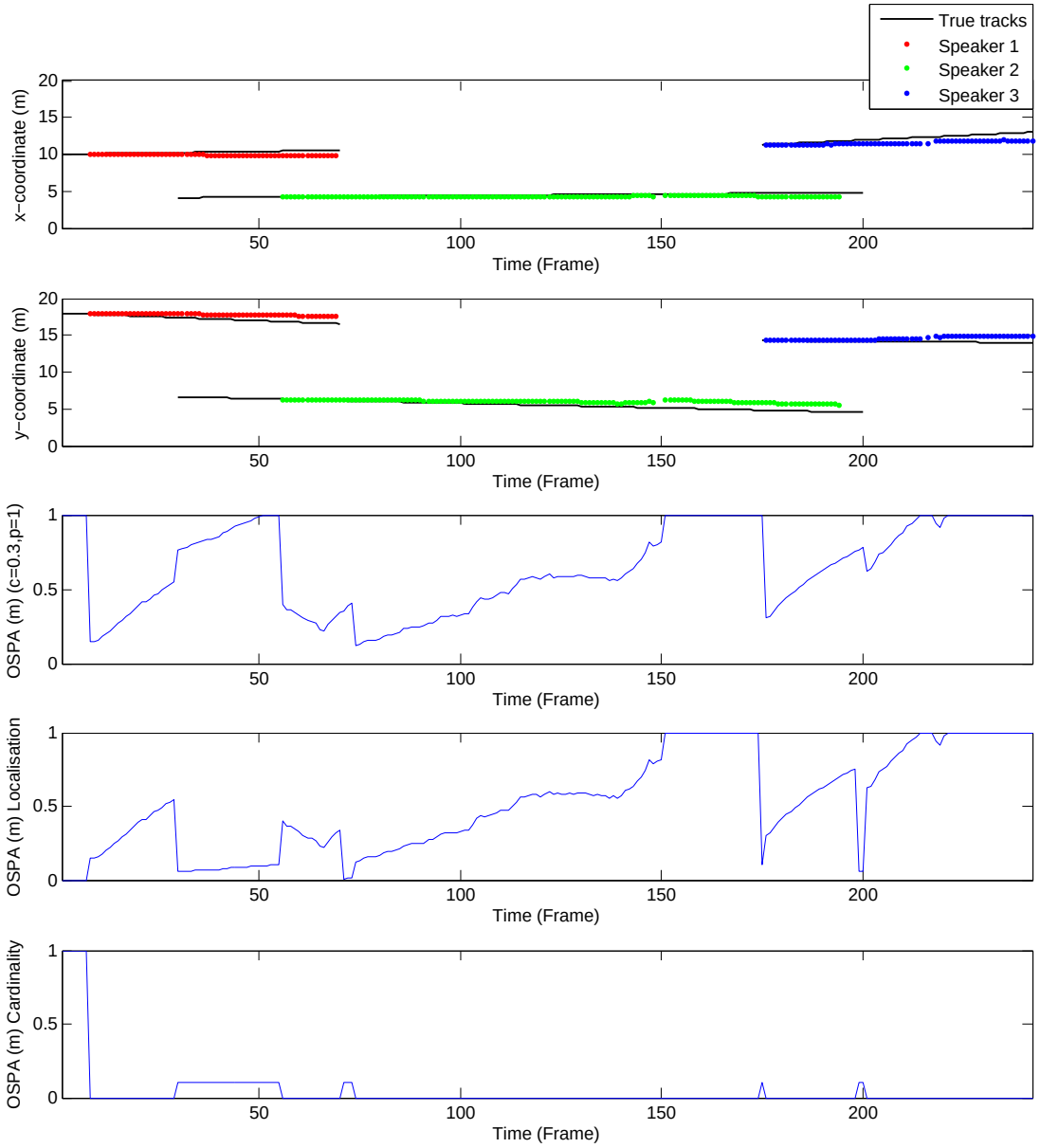


Figure 6.25: Tracking result of three speakers in a large room

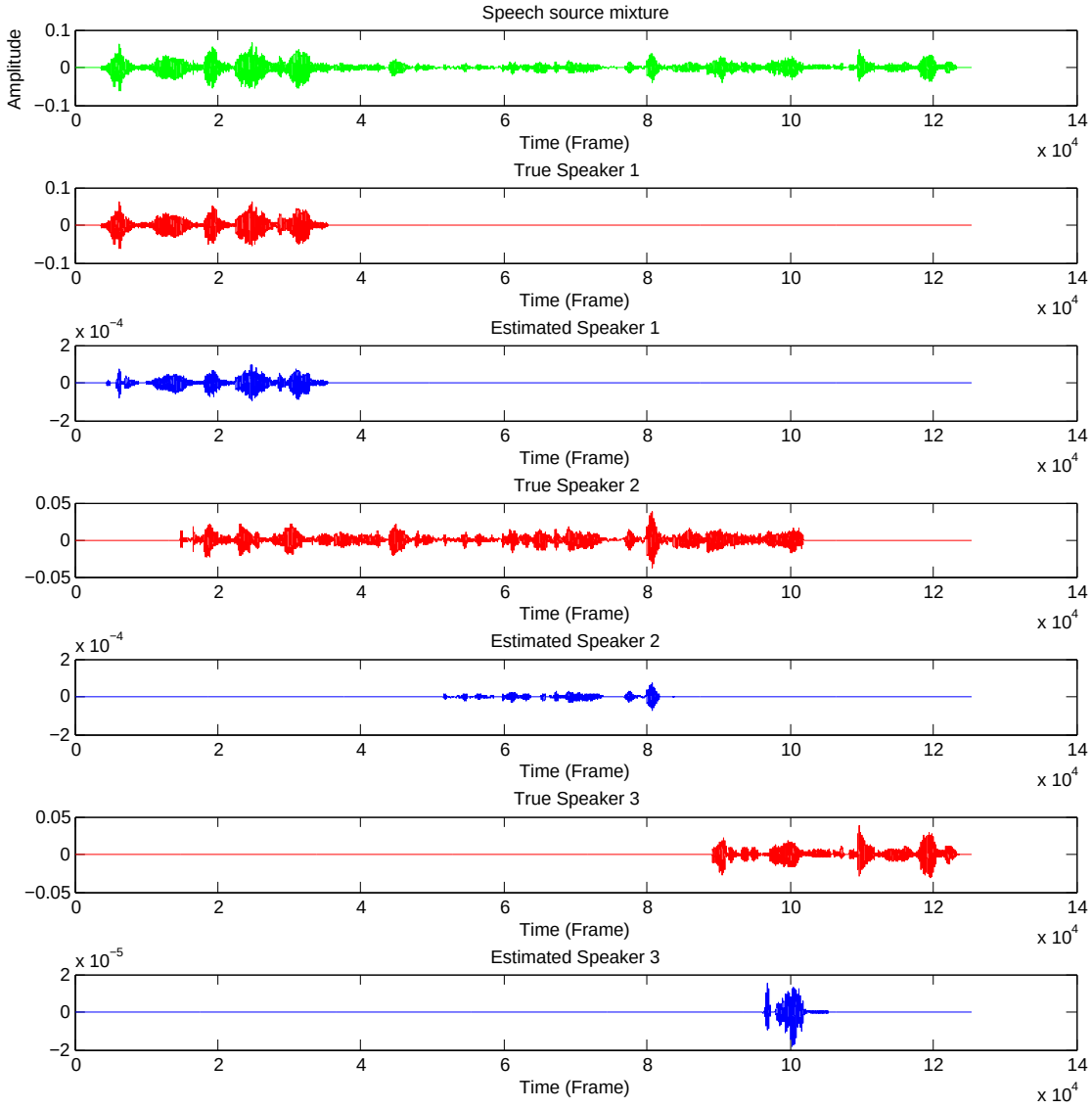


Figure 6.26: Estimated speech signals of the three speakers in small room

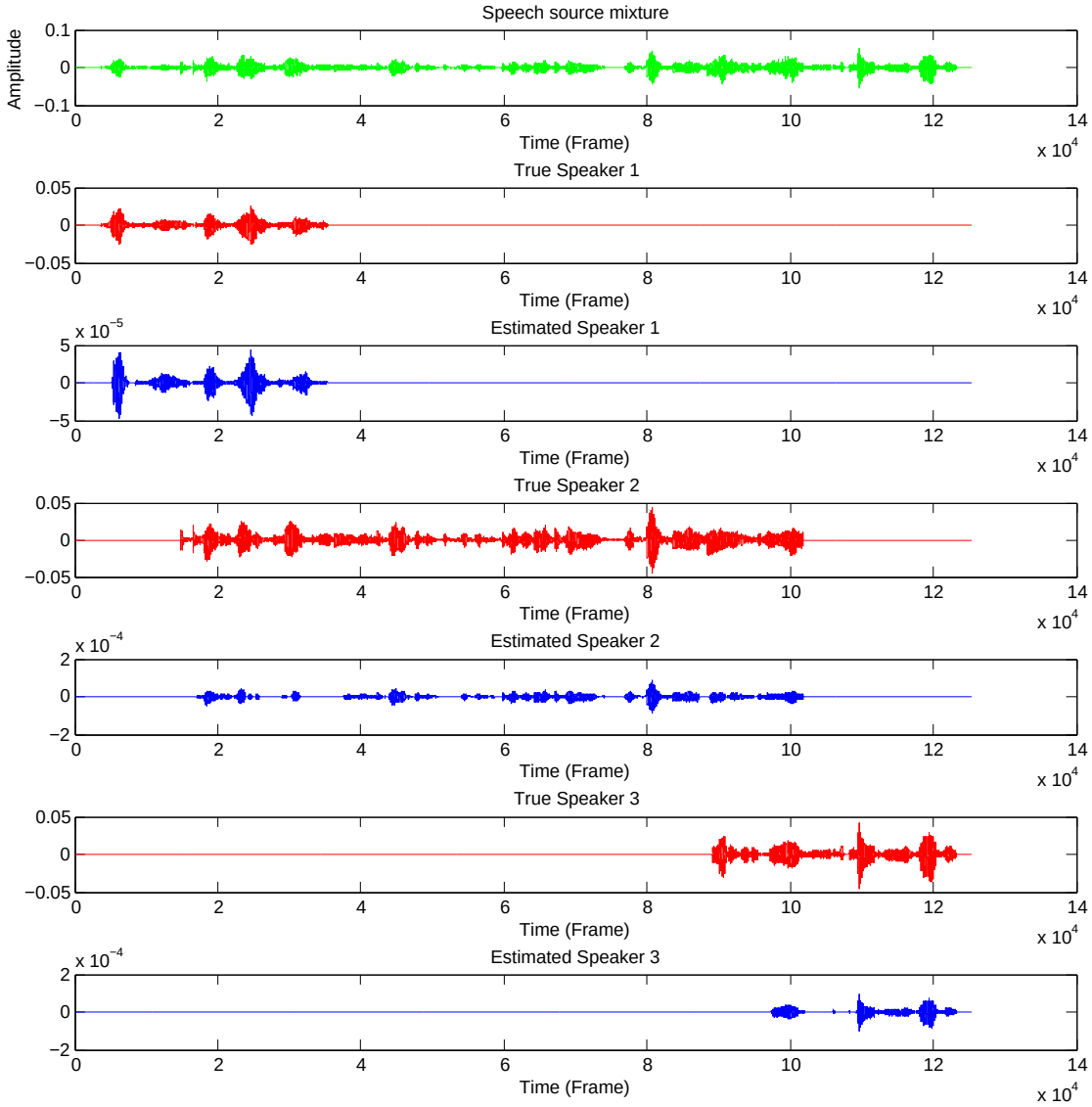


Figure 6.27: Estimated speech signals of the three speakers in medium room

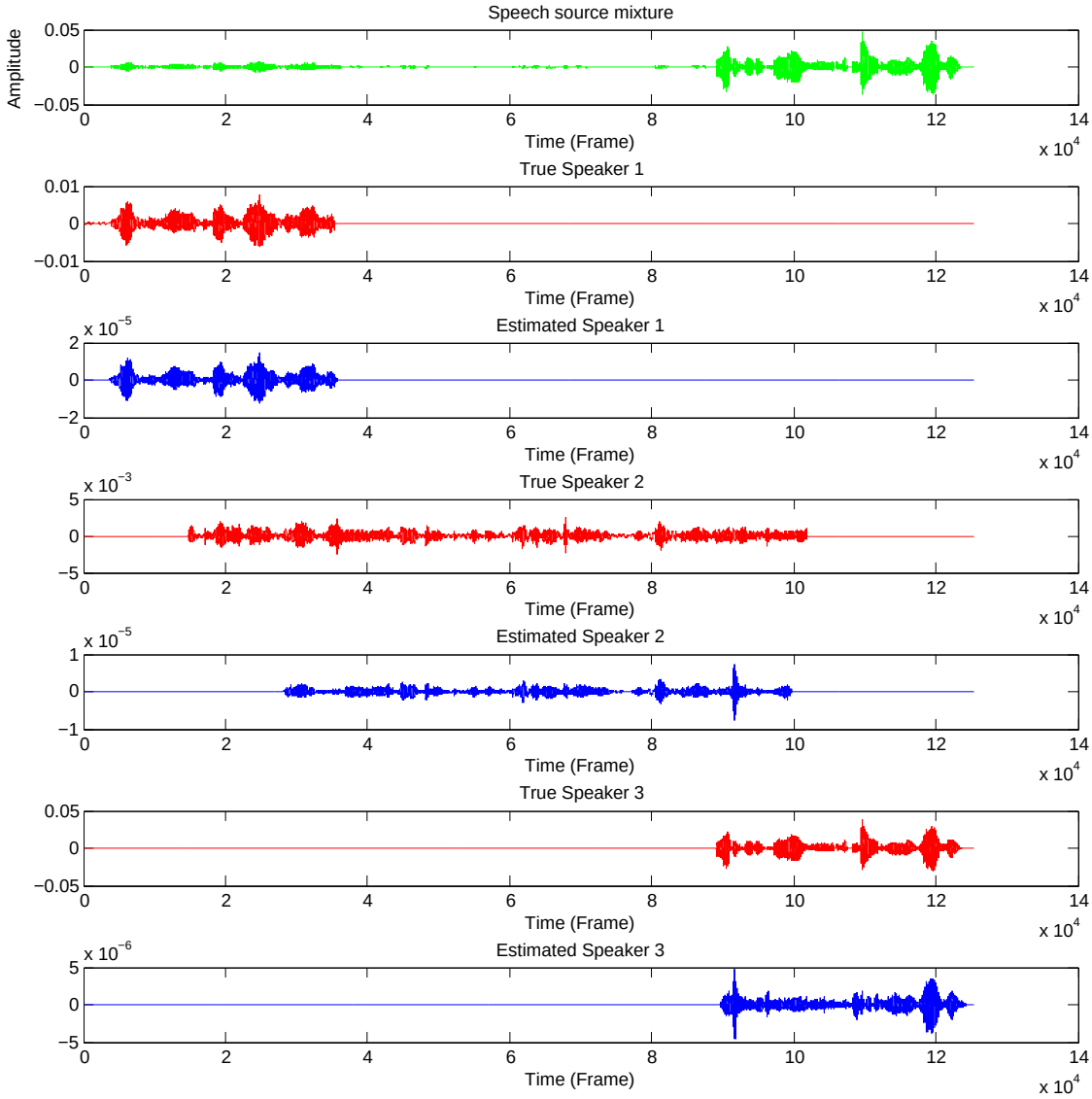


Figure 6.28: Estimated speech signals of the three speakers in large room

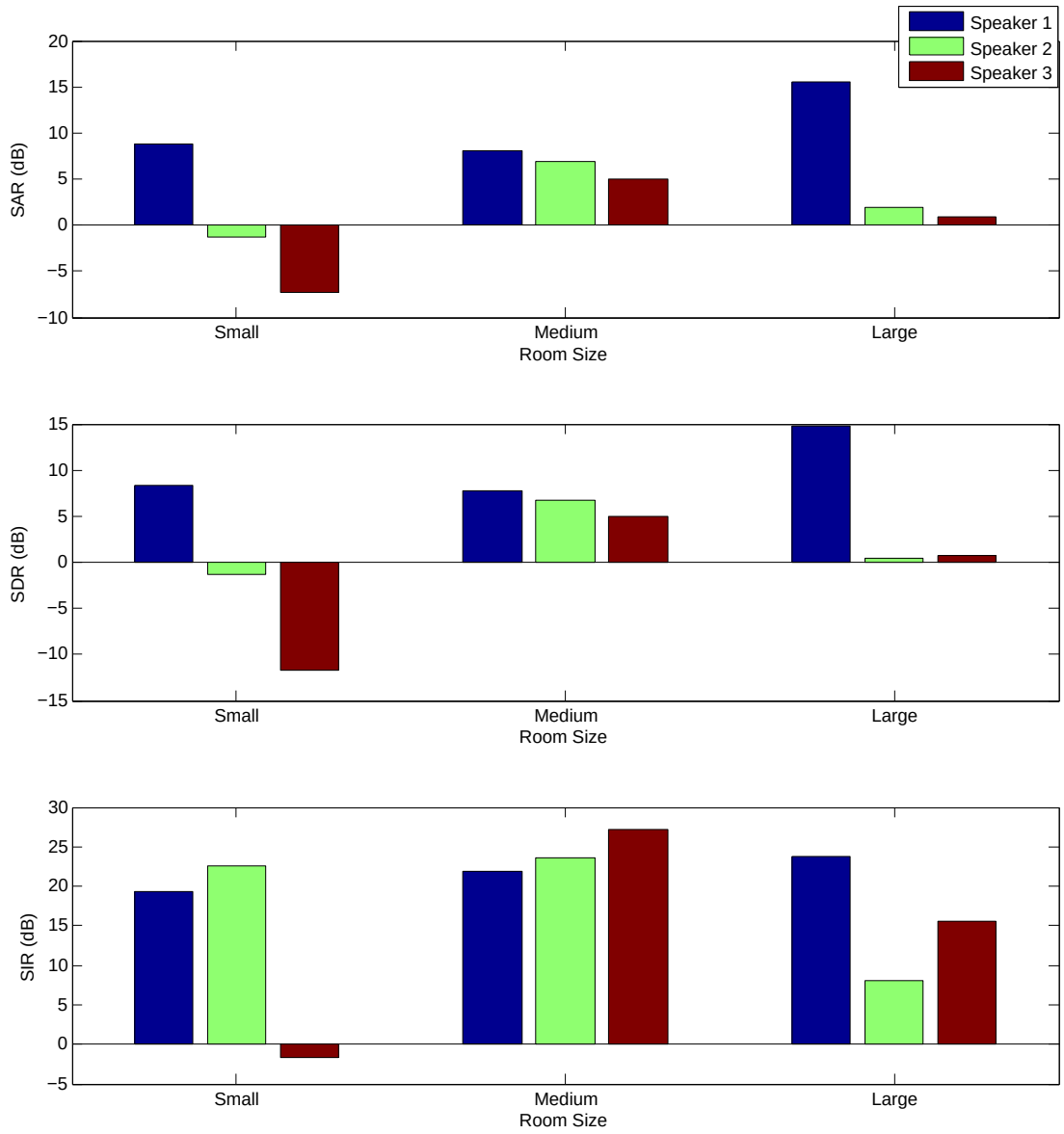


Figure 6.29: Separation performance in different room sizes

6.7.1 Discussion

The trend in figures 6.23, 6.24, 6.25, 6.26, 6.27, 6.28 and 6.29 shows that the room reverberation effect is worsened by the decrease in the size of the room. Although all three rooms were simulated with a reverberation time of $T60 = 0.15\text{s}$ and a SNR of 30dB, the tracking and separation performances are different for each room. For the largest room, the OSPA distance is an average of 0.6728m error per target. The medium room has an average target error of 0.2336m. The average error per target for the smallest room is 0.2379m. The cardinality error component of the OSPA saturates at the point in time when two speech sources overlap.

In terms of the separation results, the trend also shows a similar deterioration as the room size decreases. The average SAR of the three speakers in the small, medium and large room are 0.0720dB, 6.6298dB, and 6.0942dB respectively. The average SDR of the three rooms are -1.5365dB, 6.5166dB, and 5.3103dB. As observed in the simulations earlier, the SIR defies the SAR and SDR trends. The average SIR of the three speakers in the small, medium and large room are 13.4178dB, 24.2610dB, and 15.7490dB respectively.

The increase in the OSPA localisation error for the large room is as expected. The cause for the increase in localisation error is the small distance between the microphone pair which results in poor resolution. The distance between the microphone pair determines the resolution of the localisation. As the distance between the microphones are relatively small compared to the distance between the speech source and the microphones, any minuscule changes in the received TDoA will result in large changes in the estimated position of the speech source. This is a limitation of the design as the distance between the microphones are kept to be less than to prevent spatial aliasing. Despite this limitation, the proposed solution still managed to track the multiple speech sources without losing any part of the speech sources in the large room.

One possible reason for the deterioration of the performance in decreased room size is the diffusion of the reverberation noise. In a smaller room, the reflection of the true source is less diffused and it reaches the microphone pairs before it can be completed diffused. On the other hand, a larger space allows the reflection of the true source ample time to diffuse. Thus, the effect of reverberation is reduced in a larger space. The separation results show that there is a strong relationship between the tracking and separation performance. During instances when the speech sources fail to be tracked in the smaller room, the TF mask for the time period cannot be created. Thus, there is a silence period where a speech activity should be in the reconstructed speech. Further investigations

6.8 Effects of Hard Masking vs Soft Masking

Hard masking technique uses the original mask scoring system proposed in DUET [4]. In this proposed solution, a soft masking technique is used. The soft mask is

constructed by comparing the similarity between the estimated "true features" and the estimated "features" in each time frequency bin using a normal distributed function. In order to determine the quality of separation, two different evaluations were employed - an objective evaluation carried out using the *BSS eval Toolkit*[124] and a subjective listening test using the Mean Opinion Score (MOS) [**?** , p490-491].

In the objective evaluation, the speech source mixtures in rooms with reverberation times of $T60 = 0.05\text{s}$, $T60 = 0.15\text{s}$, $T60 = 0.25\text{s}$, $T60 = 0.35\text{s}$ and $T60 = 0.45\text{s}$ were separated using a hard masking technique and a soft masking technique. The reconstructed speech signals were then evaluated using the *BSS eval Toolkit* to compare the differences between hard masking and soft masking. The SNR of the speech signals were kept constant at 30dB. A speech source mixture in an ideal scenario with no reverberation and noise was used as a control variable for this experiment. The separation results of the soft masking technique is shown in figure 6.30 while the separation results of the hard masking technique is shown in figure 6.31.

The informal MOS test was administered by distributing surveys to 12 expert listeners. The listeners were asked to listen to reconstructed speech sources of the three speakers and to rate them from a score of 5(Excellent) to 1(Poor). The test was divided into three sections with each section featuring the reconstructed speech sources of the corresponding speaker. In order to condition the participants so they have an equal standard on the perceptual quality, samples of excellent, average and poor speech separation were played at the start of each section in the listening test. This "anchoring phase" is important as the perception of quality differs from listeners to listeners and this subjective range of differences should be minimised [**?** , p490-491]. The responses of the listeners were then consolidated to form the MOS. The results of the MOS are shown in table 6.1.

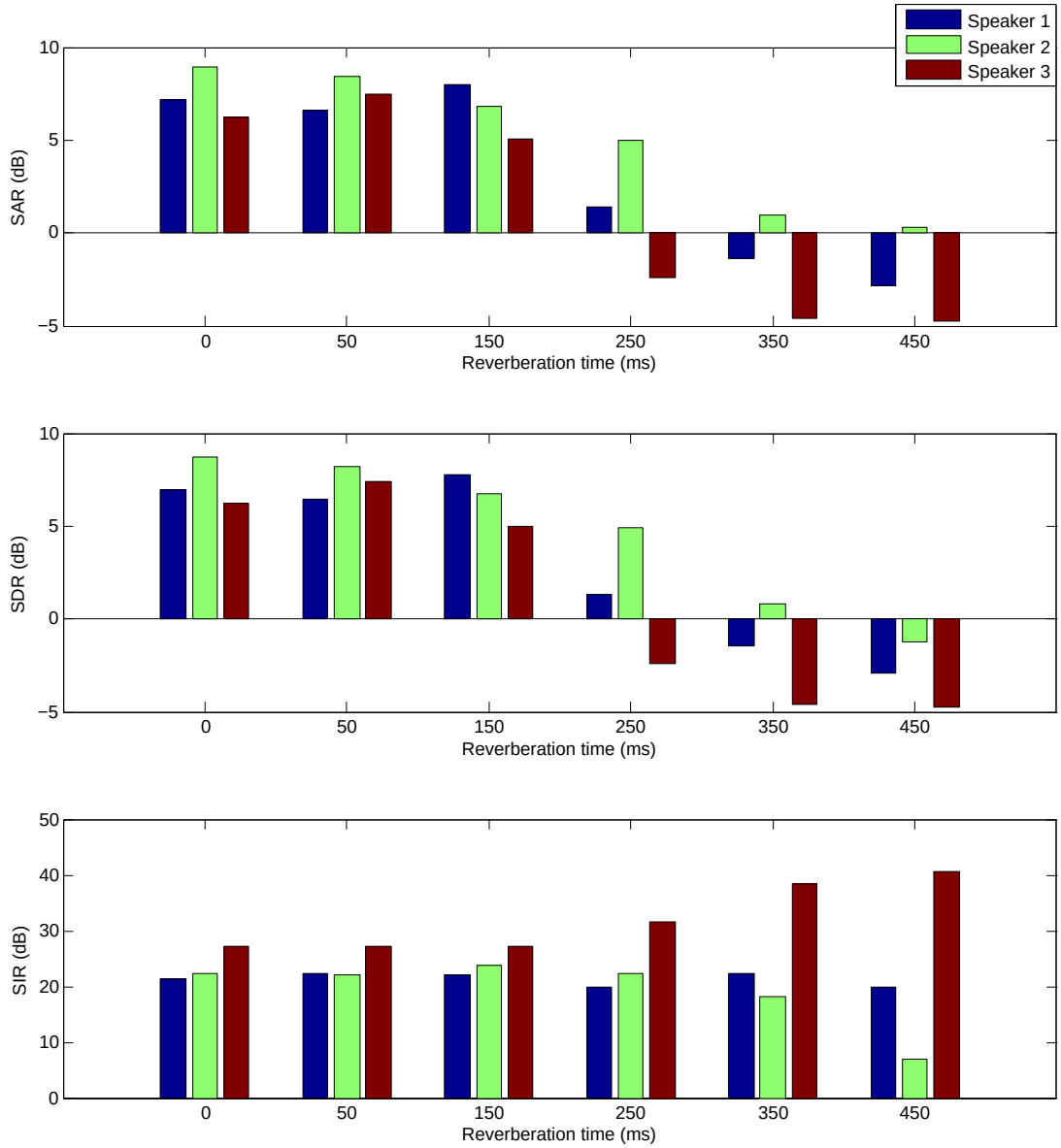


Figure 6.30: Separation performance in different room reverberations using a soft mask

Table 6.1: Mean Opinion Score for Reconstructed Speech Sources(Hard Masking vs Soft Masking)

Speakers	Mean Opinion Score (MOS)
1 (Hard Mask)	2.000
2 (Hard Mask)	2.333
3 (Hard Mask)	2.333
1 (Soft Mask)	2.833
2 (Soft Mask)	3.417
3 (Soft Mask)	3.917

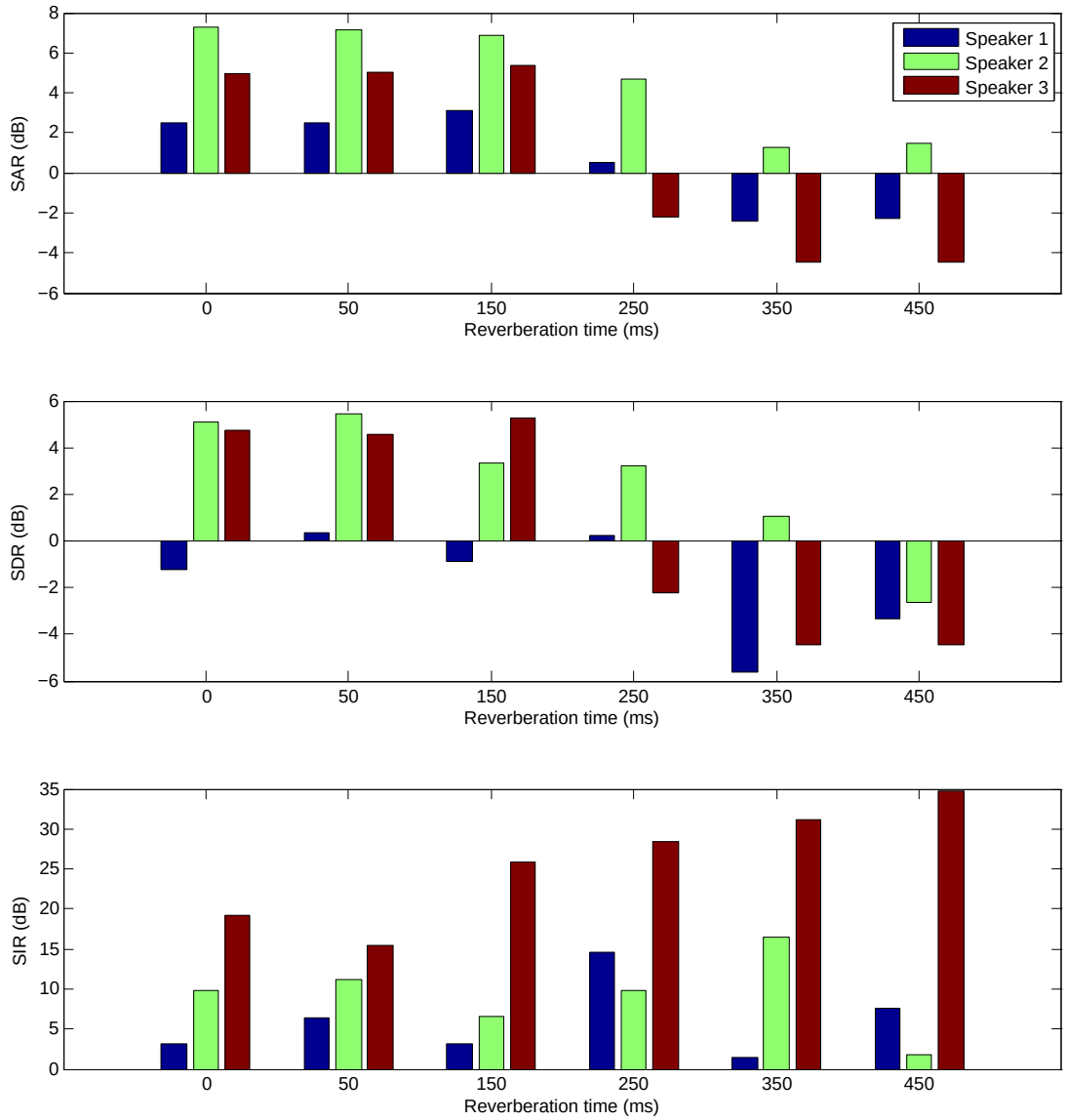


Figure 6.31: Separation performance in different room reverberations using a hard mask

6.8.1 Discussion

As shown in the results in figures 6.30 and 6.31, the soft mask proposed by this thesis has a better separation performance. The highest average SAR and SDR achieved by the soft masking technique are 7.4528dB, and 7.3027dB respectively while the highest average SAR and SIR achieved by the hard masking technique are 4.9002dB, and 2.8651dB. The soft masking technique also performed better than the hard masking technique in the other room simulations. The average SAR and SDR across all the room simulations for the soft masking technique are 3.1345dB and 2.9581dB. On the other hand, the average SAR and SDR across all the room simulations for the hard

masking technique are only 2.0521dB and 0.4739dB. Based on the MOS, there is a general consensus among the listeners that the speech sources reconstructed using a soft masking technique sound better. A main reason for this is the decrease in noise artifacts in the reconstructed speech sources using a soft masking technique. In a hard masking technique, the acoustic features are grouped into clusters using a hard threshold. The mask constructed using such a technique is a binary mask. The soft mask proposed in this thesis is a Gaussian mask which estimates the probability of the acoustic feature belonging to a particular speech source. As shown in the previous results, the performance of the speech source separation and reconstruction is closely related to the tracking performance. Hence, the soft masking technique proposed in this thesis allows for a smoother reconstruction of the speech sources during instances when the estimated acoustic features are not very accurate. In such instances, the inaccurate acoustic features still have the possibility of being assigned to their respective speech sources instead of being totally secluded. The inclusion of these acoustic features in the mask reduces the "musical noise" and produces a more natural sounding reconstructed speech source. The results of the separation performance between the hard masking and soft masking techniques which shows that soft masking produces better separation coincide with findings by other researchers such as those in [104] and [103].

6.8.2 Summary

In summary, the proposed solution is shown to be capable of tracking and separating multiple moving speech sources in various acoustic environments. In scenarios with low reverberation time and high SNR speech mixtures, the proposed solution performed excellently in terms of tracking and speech source separation. The tracking and speech separation performance deteriorates with the increase in reverberation time and noise levels although the effects of reverberation are more significant than noise. The main reason for this is the model mismatch. The TF ratio used to calculate the instantaneous attenuation and delay works under the assumption that the room reverberation is low. When the room reverberation increases, the error that is present in the calculation of these acoustic features also increases resulting in a more inaccurate observed measurement. Besides this, the model mismatch is also present in the CBMeMBer

filter when room reverberation is high. The CBMeMBer assumes that the observed measurements are independent but the noise contributed by reverberation is not really independent. The reflections of the speech sources also follow a movement pattern as the speech sources move inside the room. This results in the CBMeMBer filter mistaking the observed measurements due to reverberation for speech source targets. The effects from noise are not as significant because the behaviour of noise matches the modelling assumptions. The results in Chapter 6 also show that reverberation is exacerbated by a smaller room size. The type of masking technique used for the source separation affects the perception of the reconstructed speech sources. Speech sources reconstructed using a soft masking technique has less "musical artifact" and sound more natural.

Chapter 7

Conclusions and Future Works

From here on, the stage is mine!

– Kouta Kazuraba

7.1 Conclusion

In this thesis, a new approach to solving the source separation problem in the "conference room scenario" was proposed. The source separation was no longer viewed as a purely acoustic problem but as multi-target tracking problem as well. The proposed solution allows the moving speech sources in the room to be tracked based on their acoustic features. Identities are assigned to the tracked speech sources using a track management extension. Both the state and identity information are used to associate the acoustic features with their respective speech sources. TF masks which are applied to the speech source mixture to separate the sources are constructed using the state and identity information.

The source separation problem in the "conference room scenario" was defined in Chapter 2. A system model of the received acoustic signals was also presented in Chapter 2. Essentially, the problem is treated as both an acoustic problem as well as a multi-target tracking problem and this complex problem is broken down further into smaller problems and systematically solved for each. Hence, literature from both the acoustic and tracking field was studied in order to gain a better insight of the different approaches available as well as their limitations. A BSS approach was selected for acoustic extraction while the CBMeMBer was chosen to track the active speech sources

in the room.

In Chapter 3, the details of the BSS approach adapted for extracting the acoustic features were discussed. There are various BSS approaches which are capable of performing source separation but BSS via signal sparsity approach is selected to extract the acoustic features of the moving speech sources. This is mainly due to the fact this BSS approach is capable of extracting the acoustic features even in underdetermined scenarios whereby the number of active speech sources exceed the number of microphones available in the room. The current BSS via signal sparsity has its limitation in extracting acoustic features from moving sound sources and clustering the extracted features. Thus, a frame-by-frame analysis of the speech source mixture is proposed as the moving speech sources can be assumed to be pseudostationary in the time frame in which they are analysed. The limitation in the clustering of the extracted acoustic features from the moving speech sources is overcome by using the CBMeMBer with track management extension.

The theory behind the CBMeMBer filter and the implementation of the filter was discussed in Chapter 4. The random nature in which the active speech sources appear and disappear as well as the acoustic features generated by these speech sources mirror the classical multi-target tracking problem. Thus, it motivates the used of the CBMeMBer as part of the solution in solving the source separation problem in the "conference room scenario". The RFS models used in the CBMeMBer filter are the Poisson RFS and the multi-Bernoulli RFS. The Poisson RFS is used to represent the clutter while the multi-Bernoulli RFS is used as an approximation of the multi-target Bernoulli density. The limitation of the CBMeMBer in its application to track moving speech sources is the lack of labels for the tracked sources. The lack of an inherent labelling technique is overcome by using a post process track management extension.

In Chapter 5, the main contribution of this thesis which is to provide an online solution to the source separation problem in the "conference room scenario". This thesis examines the source separation problem in the "conference room scenario" not just as an acoustic problem but a multi-target tracking problem as well. By utilising BSS via signal sparsity to extract the acoustic features of the moving speech sources on a frame-by-frame basis, the acoustic problem is turned into a classical multi-target

tracking problem. The SMC implementation of the CBMeMBer filter is used to estimate the number of active speech sources as well as their locations in the room based on the recursive measurements of the acoustic features. In order to associate the correct acoustic features to their respective speech sources, labels are added to the tracked speech sources by a track management extension. The acoustic feature and the labels are required to construct the TF masks. A soft masking technique is applied in this thesis. After the construction of the TF masks, the speech sources are separated by applying the TF masks onto the speech mixture.

The proposed solution is evaluated in different environments in Chapter 6. The proposed solution was evaluated in different reverberation time, noise levels and room sizes. The difference in perception of the reconstructed speech sources using a hard masking technique and the proposed soft masking technique was also evaluated in Chapter 6. In the evaluation of reverberation time in a room, the proposed solution performed as expected. There is good tracking and separation performance when the room reverberation is low and this performance deteriorates with the increase in reverberation time. A similar trend is also noted when speech source mixtures with different SNRs were evaluated. As the SNR decreases, the tracking and separation performance suffers. However, the effect of noise on the performance of the solution is not as profound as the effect of reverberation. This is due to the correlation of the speech sources' reflection. The dynamic model used in the CBMeMBer assumes independence among the readings and the correlation of the reverberation affects the tracking performance of the filter. On the other hand, the noise is independent white Gaussian noise and it fits the modelling assumption in the dynamics model. Thus, the effect of noise on the tracking performance is not as serious as the effect of reverberation. In conclusion, the experiment results show that the proposed solution is a viable method of tracking and separating moving speech sources in a "conference room scenario" with low to medium reverberation.

7.2 Limitations and Future Work

Although the proposed solution is capable of tracking and separating moving speech sources in rooms with low to medium reverberation, it is not without its limitations.

The main limitation faced by the proposed online recursive solution is the approach used in the extraction of acoustic features. The proposed approach of extracting acoustic features by exploiting the signal sparsity is a good choice when it is assumed that the number of speech sources is unknown and the underdetermined scenario might arise. However, if the number of microphones is more than the maximum number of speech sources in the room, other approaches can be used. As stated in Chapter 3, the BSS via signal sparsity approach relies on the assumption of *W-disjoint orthogonality* to achieve sound source separation. *W-disjoint orthogonality* assumes that only one sound source is active at any given time frequency. This assumption holds true for speech sources as speakers usually take turns to speak in a "conference room scenario". However, this assumption is violated when the reverberation in the room is medium to high. Due to the effects of reverberation, the assumption that only one sound source is active at any given time frequency bin is no longer true as the speech sources are reflected. Hence, the acoustic features extracted using this method does deteriorate when the reverberation in a room increases. Thus, an enhanced acoustic feature extraction technique should be able to improve the tracking and separation performance of the proposed solution.

7.2.1 Enhanced Acoustic Feature Extraction

A possible method of improving the performance of the proposed is to include a Voice Activity Detector (VAD) algorithm as part of the acoustic features extraction process. It was shown in [67] that a VAD is used before a block of speech is processed by the PF in order to determine the level of speech activity in the block of data. The integration of the VAD allows the speech sources to be tracked again after a long period of silence. In [128], the incorporation of the VAD allows the localisation technique to detect a sound source with a lower gain. Hence, by pre-processing the blocks of speech signals with a VAD, only useful measurements with high levels of speech activity will be processed by the multi-target tracking algorithm. Furthermore, the VAD results can also be used to determine the probability of a speech source in the birth model of the RFS filter. This will allow the speech sources to be tracked more accurately and reliably.

In [30] and [31], CASA technique has been shown to complement BSS techniques in separating sound sources. Thus, it is possible to incorporate CASA techniques as

part of the solution in the future in order to improve the performance of this thesis' proposed solution. One of the CASA techniques considered is the neuro-fuzzy localisation technique introduced in [129]. This CASA technique has been shown to track multiple speakers in high reverberation time in [130]. Furthermore, the multi-target tracking technique used in [130] is CPHD. As it was shown to work with an RFS multi-target tracking technique, the neuro-fuzzy localisation technique can be integrated into the technique proposed by this thesis without much complications.

The enhanced acoustic feature extraction will allow for a more accurate estimation of the active speech sources in the room. This will not just improve the tracking results of the speech sources but the source separation performance as well. A more accurate estimate of the speech source location will result in a more accurate TF mask for the speech sources.

7.2.2 Inherent Labels in Multi-target Tracking Filter

In this thesis, the multi-target tracking technique used is the CBMeMBer filter. The exclusion of the labels in the CBMeMBer filter allows it to be mathematically tractable. The labels for the tracked speech sources are added post tracking process with a track management extension. This is a computationally cost effective solution for an online speech tracking technique. This implementation of labelling for the tracked speech sources has its limitations. As the track management relies on the kinematics of the speech sources in order to associate the speech sources, the identification of these speech sources after a silence period in the speech is not optimal. If the silence period is longer than the allowable missed target period set in the algorithm, the same target will be considered a new speech source. Furthermore, when the tracked speech sources are close to each other, the measurements will only confirm for one of the speech sources as the CBMeMBer assumes independence in the measurement of the tracked targets [13]. As a result of this, a multi-target tracking technique which performs labelling as part of the Bayesian recursion is a will be appropriate technique if tracking accuracy is the main concern instead of computational cost.

The Generalised Labelled Multi-Bernoulli (GLMB) filter [131] is an RFS filter which was developed after the CBMeMBer. Unlike the CBMeMBer, the GLMB filter incorporates a unique label in the state vector for the tracked targets. The GLMB filter

is able to capture in inter-dependence between the tracked targets [131]. This allows targets which are close to each other to be tracked without loss of targets. The viability of the GLMB in tracking closely spaced targets are shown in [131]. By applying the GLMB filter to track speech sources, the speakers can be distinguished even if they move close to each other. As the labels are inherently incorporated as part of the state vectors, it is expected that the speech sources will continuously be tracked even after a gap or a long pause in the speech. This hypothesis will be studied in future work which uses the GLMB filter to track instead of the CBMeMBer.

7.2.3 Measurement driven birth model

Another method of improving the tracking performance is to use an adaptive birth model [132]. In a more realistic "conference room scenario", the locations whereby the targets appear are unknown *a priori*. The constant birth model has to cover the entire room if the all the targets are to be tracked as targets which appear outside the birth model do not gets confirmed as tracks. Instead of increasing the covariance of the birth model so it covers the whole room, a measurement updated birth model is a more efficient technique of confirming targets based on the measurements received. Speech source tracking and identification after a long silence period is very challenging. A reason for this is the location in which the speech source reappear is not necessarily within the boundaries of the birth model. With the use of a measurement driven birth model, the effect of the silence period on the tracking performance can be mitigated. As the location of birth model is updated according to the received measurements, the updated birth model will be closer to the speech source's last location before the silence period. Hence, there is less chance that the speech source will exceed the boundary of the birth model and a higher chance of recovery after the silence period.

7.2.4 Microphone Placement and Selection

In this thesis, multiple microphones are used to track and separate the speech sources in the "conference room scenario". The readings from multiple pairs of microphone allows the speech sources to be tracked in two dimension (Cartesian Coordinates). However, the effects of microphone placement and the order in which the microphone measurements are selected are not studied in this thesis. In reality, the power of the

signal received by the microphones changes according to the location of the speech sources. The power is expected to be higher if the source is closer to a particular microphone. This information can be exploited as the microphones which are closer to the tracked speech sources should be selected first for the update. This will allow for more reliable information to be used as measurements for the update.

Another reason to have a smart microphone selection algorithm is to prevent target occlusion during the measurements. In a scenario whereby the tracked speech sources fall on the same hyperbolic TDoA line of uncertainty, certain microphone pairs will only have a true measurement as both objects share the same TDoA. In such a scenario, only an object will be confirmed as a target if the measurements from these microphone pairs are used to update the tracks. A more efficient method is to determine the number of targets picked up by the microphone pairs and update the measurements starting with the microphone pairs which pick up the most targets in the previous time frame. Thus, the loss of targets due to measurement occlusion can be minimised.

During speech source separation stage, an efficient selection method for the microphones to be used will result in higher quality separated speech. As discussed earlier, the power received by each microphone will be different depending on the location of the speech sources. An improvement in the quality of the reconstructed speech is expected if the TF masks are applied onto signals from microphones which are closest to the speech sources at each given time frame.

References

[1] J. Huang, N. Ohnishi, and N. Sugie, "A biomimetic system for localization and separation of multiple sound sources," *Instrumentation and Measurement, IEEE Transactions on*, vol. 44, no. 3, pp. 733–738, 1995.

[2] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual review of psychology*, vol. 42, no. 1, pp. 135–159, 1991.

[3] J.-F. Cardoso, "Source separation using higher order moments," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 2109–2112.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[5] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[6] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *Signal Processing, IEEE Transactions on*, vol. 45, no. 2, pp. 434–444, 1997.

[7] R. M. Karp, "On-line algorithms versus off-line algorithms: How much is it worth to know the future?" in *International Federation for Information Processing (IFIP) Congress (1)*, vol. 12, 1992, pp. 416–429.

[8] S. Albers, "Online algorithms: a survey," *Mathematical Programming*, vol. 97, no. 1-2, pp. 3–26, 2003.

[9] ——, "Online algorithms," in *Interactive Computation*. Springer, 2006, pp. 143–164.

[10] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013, vol. 3.

[11] X. Bian, G. D. Abowd, and J. M. Rehg, "Using sound source localization in a home environment," in *Pervasive Computing*. Springer, 2005, pp. 19–36.

[12] R. Mahler, "Multitarget bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, 2003.

[13] B.-T. Vo, B.-N. Vo, and A. Cantoni, "The cardinality balanced multi-target multi-bernoulli filter and its implementations," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 409–423, 2009.

[14] N. Grbic, X. Tao, S. Nordholm, and I. Claesson, "Blind signal separation using overcomplete subband representation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 524–533, 2001.

[15] M. Pedersen, J. Larsen, U. Kjems, and L. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook*, pp. 1065–1084, 2007.

[16] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 88, no. 7, pp. 1640–1655, 2005.

[17] C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal processing*, vol. 24, no. 1, pp. 1–10, 1991.

[18] A. Hyvärinen, "Independent component analysis by minimization of mutual information," Helsinki University of Technology, Dept of Computer Science and Engineering, Laboratory of Computer and Information Science, Rakentajanaukio 2 C, FIN-02150 Espoo, Finland, Tech. Rep., August 1997.

[19] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[20] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[21] S. Ikeda and N. Murata, "A method of ica in time-frequency domain," in *in Proc. ICA*. Citeseer, 1999.

[22] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Solving the permutation problem of frequency-domain bss when spatial aliasing occurs with wide sensor spacing," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.

[23] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on*, vol. 2. IEEE, 2003, pp. 411–414.

[24] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007.

[25] S. Low, "Speech enhancement using microphone arrays," Ph.D. dissertation, Dept of Electrical and Computer Science Engineering, Curtin University, Kent St, Bentley WA 6102 Australia, May 2005.

[26] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 405–413, 1993.

[27] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[28] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[29] A. van Der Kouwe, D. Wang, and G. Brown, "A comparison of auditory and blind separation techniques for speech segregation," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 189–195, 2001.

[30] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 1, pp. 6–6, 2007.

[31] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source–filter-based single-channel speech separation using pitch information," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 242–255, 2011.

[32] K. Nakadai, H. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition," in *Seventh International Conference on Spoken Language Processing*, 2002.

[33] H. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational auditory scene analysis and its application to robot audition," in *Informatics Research for Development of Knowledge Society Infrastructure, 2004. ICKS 2004. International Conference on*. IEEE, 2004, pp. 73–80.

[34] D. Wang and G. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[35] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 1589–1592.

[36] S. Rickard, "The DUET blind source separation algorithm," *Blind Speech Separation*, pp. 217–241, 2007.

[37] A. Koutvas, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2000. ICASSP 2000.*, vol. 2. IEEE, 2000, pp. II1133–II1136.

[38] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Realtime blind source separation for moving speakers using blockwise ica and residual crosstalk-subtraction," in *Proceedings of International Symposium on Independent Component Analysis (ICA'03)*, April 2003, pp. 975–980.

[39] B. Loesch and B. Yang, "Online blind source separation based on time-frequency sparseness," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 117–120.

[40] ——, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. IWAENC*, 2008.

[41] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Springer, 2001, pp. 157–180.

[42] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[43] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, 1986.

[44] B. Kwon, Y. Park, and Y. Park, "Analysis of the gcc-phat technique for multiple sources," in *Control Automation and Systems (ICCAS), 2010 International Conference on*. IEEE, 2010, pp. 2070–2073.

[45] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech & Language*, vol. 9, no. 2, pp. 153–169, 1995.

[46] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[47] P. R. Roth, "Effective measurements using digital signal analysis," *Spectrum*, 1971.

[48] A. Karbasi and A. Sugiyama, "A new doa estimation method using a circular microphone array," in *Proc. Euro. signal process. conf.(EUSIPCO)*, 2007, pp. 778–782.

[49] G. Doblinger, *Localization and tracking of acoustical sources*. Springer, 2006.

[50] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.* IEEE, 2008, pp. 4349–4352.

[51] C. Segura, A. Abad, J. Hernando, and C. Nadeu, "Speaker orientation estimation based on hybridation of gcc-phat and hlbr," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[52] H. L. Van Trees, *Part IV Detection, Estimation, and Modulation Theory, Optimum Array Processing*. John Wiley & Sons, 2002.

[53] M. S. Fidan, "Multiple signal classification method in direction of arrival estimation," *Electrical Electronics Engineering Department*, pp. 2–6, 2007.

[54] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8. IEEE, 1983, pp. 336–339.

[55] L. C. Godara, "Application of antenna arrays to mobile communications. ii. beam-forming and direction-of-arrival considerations," *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1195–1245, 1997.

[56] J. Krolik and D. Swingler, "Focused wide-band array processing by spatial resampling," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 2, pp. 356–360, 1990.

[57] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 4, pp. 823–831, 1985.

[58] A. Vural, "Effects of perturbations on the performance of optimum/adaptive arrays," *Aerospace and Electronic Systems, IEEE Transactions on*, no. 1, pp. 76–87, 1979.

[59] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[60] J. Griffiths, "Adaptive array processing. a tutorial," *Communications, Radar and Signal Processing, IEE Proceedings F*, vol. 130, no. 1, p. 3, 1983.

[61] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 10, pp. 1365–1376, 1987.

[62] Y. Zhao, X. Chen, and B. Wang, "Real-time sound source localization using hybrid framework," *Applied Acoustics*, vol. 74, no. 12, pp. 1367–1373, 2013.

[63] A. Johansson and S. Nordholm, "Robust acoustic direction of arrival estimation using root-srp-phat, a realtime implementation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05).*, vol. 4. IEEE, 2005, pp. iv–933.

[64] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[65] B.-N. Vo, S. Singh, and W. Ma, "Tracking multiple speakers using random sets," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2004. ICASSP 2004.*, vol. 2. IEEE, 2004, pp. ii–357.

[66] P. Pertilä, "Online blind speech separation using multiple acoustic speaker tracking and time–frequency masking," *Computer Speech & Language*, vol. 27, no. 3, pp. 683–702, 2013.

[67] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 28–28, 2007.

[68] F. Antonacci, D. Riva, D. Saiu, A. Sarti, M. Tagliasacchi, and S. Tubaro, "Tracking multiple acoustic sources using particle filtering," in *Eur. Signal Processing Conf.(EUSIPCO)*, 2006.

[69] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[70] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying rao-blackwellised monte carlo data association," in *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*. IEEE, 2009, pp. 253–256.

[71] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.

[72] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174–188, 2002.

[73] M. Mallick, V. Krishnamurthy, and B.-N. Vo, *Integrated tracking, classification, and sensor management: theory and applications*. John Wiley & Sons, 2012.

[74] E. A. Lehmann, A. M. Johansson, and S. Nordholm, "Modeling of motion dynamics and its influence on the performance of a particle filter for acoustic speaker tracking," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*. IEEE, 2007, pp. 98–101.

[75] E. Lehmann, "Particle filtering methods for acoustic source localisation and tracking," Ph.D. dissertation, The Australian National University, 2004.

[76] A. Piersol, "Time delay estimation using phase data," *IEEE Transactions on Acoustics, speech and signal processing*, vol. 29, no. 3, pp. 471–477, 1981.

[77] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using tdoa measurements: a random finite set approach," *Signal Processing, IEEE Transactions on*, vol. 54, no. 9, pp. 3291–3304, 2006.

[78] D. E. Sturim, M. S. Brandstein, and H. F. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 1. IEEE, 1997, pp. 371–374.

[79] M. Swartling, N. Grbic, and I. Claesson, "Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 4. IEEE, 2006, pp. IV–IV.

[80] D. B. Reid, "An algorithm for tracking multiple targets," *Automatic Control, IEEE Transactions on*, vol. 24, no. 6, pp. 843–854, 1979.

[81] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *Oceanic Engineering, IEEE Journal of*, vol. 8, no. 3, pp. 173–184, 1983.

[82] L. Lin, Y. Bar-Shalom, and T. Kirubarajan, "Track labeling and phd filter for multitarget tracking," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 42, no. 3, pp. 778–795, 2006.

[83] M. Rutten, J. Williams, N. Gordon, J. Stauch, J. Baldwin, and M. Jah, "A comparison of jpda and belief propagation for data association in ssa," in *15th Advanced Maui Optical and Space Surveillance Technologies Conference (09 Sep 2014-12 Sep 2014: Maui, Hawaii)*, 2014.

[84] Y. Cai, "Maintaining accurate multi-target tracking under frequent occlusion," in *European Conference on Computer Vision*, 2004.

[85] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *Signal Processing, IEEE Transactions on*, vol. 59, no. 7, pp. 3452–3457, 2011.

[86] B.-N. Vo, B.-T. Vo, N.-T. Pham, and D. Suter, "Joint detection and estimation of multiple objects from image observations," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5129–5141, 2010.

[87] B.-N. Vo, S. Singh, and A. Doucet, "Sequential monte carlo methods for multitarget filtering with random finite sets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1224–1245, 2005.

[88] N. Pham, W. Huang, and S. Ong, "Tracking multiple speakers using CPHD filter," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 529–532.

[89] R. Mahler, "Phd filters of higher order in target number," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 4, pp. 1523–1543, 2007.

[90] B.-T. Vo, B.-N. Vo, and A. Cantoni, "Analytic implementations of the cardinalized probability hypothesis density filter," *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3553–3567, 2007.

[91] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential monte carlo fusion of sound and vision for speaker tracking," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 741–746.

[92] G. Lathoud, "Spatio-temporal analysis of spontaneous speech with microphone arrays," 2007.

[93] S. Naqvi, M. Yu, and J. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895–910, 2010.

[94] P. Pertila and M. Hamalainen, "A track before detect approach for sequential bayesian tracking of multiple speech sources," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4974–4977.

[95] L.-H. Kim, I. Tashev, and A. Acero, "Reverberated speech signal separation based on regularized subband feedforward ica and instantaneous direction of arrival," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2678–2681.

[96] X. Wang, S. Challa, and R. Evans, "Gating techniques for maneuvering target tracking in clutter," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 38, no. 3, pp. 1087–1097, 2002.

[97] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Doa estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *Journal of Signal Processing Systems*, vol. 63, no. 3, pp. 265–275, 2011.

[98] M. Joho, H. Mathis, and R. H. Lambert, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," in *Proceedings International Conference on Independent Component Analysis and Blind Signal Separation. Helsinki, Finland*, 2000, pp. 81–86.

[99] S. Rickard, T. Melia, and C. Fearon, "Desprit-histogram based blind source separation of more sources than sensors using subspace methods," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*. IEEE, 2005, pp. 5–8.

[100] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using desprit," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 90–90, 2007.

[101] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[102] M. Kuhne, R. Togneri, and S. Nordholm, "A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation," *Signal Processing*, vol. 90, no. 2, pp. 653–669, 2010.

[103] I. Jafari, S. Haque, R. Togneri, and S. Nordholm, "Underdetermined blind source separation with fuzzy clustering for arbitrarily arranged sensors." in *Proceedings of INTERSPEECH, 2011. INTERSPEECH 2011.*, 2011, pp. 1753–1756.

[104] V. Reju, S. Koh, and Y. Soon, "Underdetermined convolutive blind source separation via time–frequency masking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 101–116, 2010.

[105] I. Gath and A. Geva, "Unsupervised optimal fuzzy clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, no. 7, pp. 773–780, 1989.

[106] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[107] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000*, vol. 5. IEEE, 2000, pp. 2985–2988.

[108] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–529.

[109] Y. Bar-Shalom, *Tracking and data association*. Academic Press Professional, Inc., 1987.

[110] R. Mahler, *Statistical multisource-multitarget information fusion*. Norwood: Artech House, 2007.

[111] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of Nonlinear Filtering*, vol. 12, pp. 656–704, 2009.

[112] R. Douc and O. Cappé, "Comparison of resampling schemes for particle filtering," in *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*. IEEE, 2005, pp. 64–69.

[113] R. Hoseinnezhad, B.-N. Vo, B.-T. Vo, and D. Suter, "Visual tracking of numerous targets via multi-bernoulli filtering of image data," *Pattern Recognition*, vol. 45, no. 10, pp. 3625–3635, 2012.

[114] S. Wong, B.-T. Vo, B.-N. Vo, and R. Hoseinnezhad, "Multi-bernoulli based track-before-detect with road constraints," in *15th International Conference on Information Fusion (FUSION), 2012*. IEEE, 2012, pp. 840–846.

[115] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[116] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[117] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function gsc and postfiltering," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 6, pp. 561–571, 2004.

[118] M. Brandstein, J. Adcock, and H. Silverman, "Microphone-array localization error estimation with application to sensor placement," *The Journal of the Acoustical Society of America*, vol. 99, pp. 3807–3816, 1996.

[119] S. Blackrnan and A. House, "Design and analysis of modern tracking systems," *Boston, MA: Artech House*, 1999.

[120] R. Mahler, "Approximate multisensor CPHD and PHD filters," in *13th Conference on Information Fusion 2010 (FUSION'10)*. IEEE, 2010, pp. 1–8.

[121] K. Shafique and M. Shah, "A noniterative greedy algorithm for multiframe point correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 51–65, 2005.

[122] N. Chong, S. Wong, S. Nordholm, and I. Murray, "Multiple sound source tracking and identification via degenerate unmixing estimation technique and cardinality balanced multi-target multi-bernoulli filter (duet-cbmember) with track management," in *Asia-Pacific Signal and Information Processing Association, 2014. APSIPA 2014. Annual Summit and Conference*, Dec 2014, pp. 1–5.

[123] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008.

[124] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[125] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.

[126] N. Chong, S. Wong, B.-T. Vo, S. Nordholm, and I. Murray, "Multiple sound source tracking via degenerate unmixing estimation technique and cardinality balanced multi-target multi-bernoulli filter (DUET-CBMeMBer) with track management," in *IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing, 2014. ISSNIP 2014.* IEEE, 2014, pp. 1–6.

[127] E. Lehmann, "Image-source method: Matlab code implementation," http://www.eric-lehmann.com, March 2012.

[128] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*. IEEE, 2011, pp. 92–97.

[129] A. Plinge, M. H. Hennecke, and G. A. Fink, "Robust neuro-fuzzy speaker localization using a circular microphone array," in *Proc. Int. Workshop on Acoustic Echo and Noise Control, Tel Aviv, Israel*. Citeseer, 2010.

[130] A. Plinge, D. Hauschildt, M. H. Hennecke, G. Fink *et al.*, "Multiple speaker tracking using a microphone array by combining auditory processing and a gaussian mixture cardinalized probability hypothesis density filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*. IEEE, 2011, pp. 2476–2479.

[131] F. Papi, B.-N. Vo, B.-T. Vo, C. Fantacci, and M. Beard, "Generalized labeled multi-bernoulli approximation of multi-object densities," *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5487–5497, 2015.

[132] B. Ristic, D. Clark, B.-N. Vo, and B.-T. Vo, "Adaptive target birth intensity for phd and cphd filters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 2, pp. 1656–1668, 2012.